The American Economic Review

ARTICLES

98

391
601

JOSEPH A. PECHMAN
The Future of the Income Tax

KENNETH ROGOFF
Equilibrium Political Budget Cycles

GUIDO TABELLINI AND ALBERTO ALESINA
Voting on the Budget Deficit

MICHELLE R. GARFINKEL
Arming as a Strategic Investment in a Cooperative Equilibrium

ARTHUR J. ROBSON
Stackelberg and Marshall

WILLIAM PAUL ROGERSON
Quality versus Quantity in Military Procurement

DAVID SCHARFSTEIN AND PATRICK BOLTON
A Theory of Predation Based on Agency Problems in Financial Contracting

JOSEPH FARRELL AND CARL SHAPIRO
Horizontal Mergers: An Equilibrium Analysis

JANUSZ A. ORDOVER, GARTH SALONER, AND STEVEN C. SALOP
Equilibrium Vertical Foreclosure

ERIC MASKIN AND DAVID NEWBERY
Disadvantageous Oil Tariffs and Dynamic Consistency

LEONARDO AUERNHEIMER AND GABRIEL A. LOZADA
On the Treatment of Anticipated Shocks in Models of Optimal Control with Rational Expectations: An Economic Interpretation

RUSSELL W. COOPER AND JOHN C. HALTIWANGER
Inventories and the Propagation of Sectoral Shocks

RAYMOND D. SAUER AND KEITH B. LEFFLER
Did the Federal Trade Commission's Advertising Substantiation Program Promote More Credible Advertising?

AMOS TVERSKY, PAUL SLOVIC, AND DANIEL KAHNEMAN
The Causes of Preference Reversal

SHORTER PAPERS: R. W. Cooper, D. V. DeJong, R. Forsythe, and T. W. Ross; J. B. Van Huyck, R. C. Battalio, and R. O. Beil; E. Dekel-Tabak and S. Scotchmer; H. P. Young; K. R. Ihlanfeldt and D. L. Sjoquist; D. A. Barbezat and J. W. Hughes; W. E. Even; P. Kuhn; G. Jasso and M. R. Rosenzweig; G. J. Borjas; W. A. Bomberger.

CENTRAL LIBRARY

MARCH 1990

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President
GERARD DEBREU
University of California-Berkeley

President-elect
THOMAS C. SCHELLING
Harvard University

Vice-Presidents
RUDIGER W. DORNBUSCH
Massachusetts Institute of Technology
ALLAN H. MELTZER
Carnegie Mellon University

Secretary-Treasurer
C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review
ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature
JOHN PENCAVEL
Stanford University

Editor of The Journal of Economic Perspectives
JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
GEORGE A. AKERLOF
University of California-Berkeley
ISABEL V. SAWHILL
Urban Institute
STANLEY FISCHER
World Bank
LAWRENCE H. SUMMERS
Harvard University
GREGORY C. CHOW
Princeton University
SUSAN ROSE-ACKERMAN
Yale University

EX OFFICIO Member
ROBERT EISNER
Northwestern University

●Printed at Banta Company, Menasha, Wisconsin.

●Copyright © American Economic Association 1990. All rights reserved.

●No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212–2786. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002–8282), March 1990, Vol. 80, No. 1, is published five times a year (March, May, June, September, December) by the American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212–2786. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: $42.00, $50.40, or $58.80 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add $16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 1313 21st Avenue South, Suite 809, Nashville, TN 37212–2786.

THE AMERICAN ECONOMIC REVIEW

Editor

ORLEY ASHENFELTER

Co-Editors
ROBERT H. HAVEMAN
BENNETT T. McCALLUM
PAUL R. MILGROM

Production Editor
CLAIRE H. COMISKEY

Board of Editors

GEORGE A. AKERLOF
JAMES E. ANDERSON
TIMOTHY F. BRESNAHAN
HENRY S. FARBER
MARJORIE A. FLAVIN
ROBERT P. FLOOD
CLAUDIA D. GOLDIN
JO ANNA GRAY
REUBEN GRONAU
DANIEL S. HAMERMESH
KEVIN D. HOOVER
KENNETH L. JUDD
JOHN H. KAGEL
JOHN F. KENNAN
DALE T. MORTENSEN
MAURICE OBSTFELD
EDGAR O. OLSEN
JOHN G. RILEY
RICHARD ROLL
THOMAS ROMER
DAVID E. M. SAPPINGTON
KENNETH J. SINGLETON
ROBERT S. SMITH
BARBARA J. SPENCER
RICHARD TRESCH
KENNETH WEST
JOHN D. WILSON
LESLIE YOUNG

March 1990

VOLUME 80, NUMBER 1

391
1 001

Articles

The Future of the Income Tax *Joseph A. Pechman* 1

Equilibrium Political Budget Cycles *Kenneth Rogoff* 21

Voting on the Budget Deficit *Guido Tabellini and Alberto Alesina* 37

Arming as a Strategic Investment in a Cooperative Equilibrium *Michelle R. Garfinkel* 50

Stackelberg and Marshall *Arthur J. Robson* 69

Quality versus Quantity in Military Procurement *William Paul Rogerson* 83

A Theory of Predation Based on Agency Problems in Financial Contracting *David Scharfstein and Patrick Bolton* 93

Horizontal Mergers: An Equilibrium Analysis *Joseph Farrell and Carl Shapiro* 107

Equilibrium Vertical Foreclosure *Janusz A. Ordover, Garth Saloner, and Steven C. Salop* 127

Disadvantageous Oil Tariffs and Dynamic Consistency *Eric Maskin and David Newbery* 143

On the Treatment of Anticipated Shocks in Models of Optimal Control with Rational Expectations: An Economic Interpretation *Leonardo Auernheimer and Gabriel A. Lozada* 157

Inventories and the Propagation of Sectoral Shocks *Russell W. Cooper and John C. Haltiwanger* 170

Did the Federal Trade Commission's Advertising Substantiation Program Promote More Credible Advertising? *Raymond D. Sauer and Keith B. Leffler* 191

The Causes of Preference Reversal *Amos Tversky, Paul Slovic, and Daniel Kahneman* 204

Shorter Papers

Selection Criteria in Coordination Games: Some Experimental Results
Russell W. Cooper, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross 218
Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure
John B. Van Huyck, Raymond C. Battalio, and Richard O. Beil 234
Collusion Through Insurance: Sharing the Costs of Oil Spill Cleanups
Eddie Dekel-Tabak and Suzanne Scotchmer 249
Equal Sacrifice and Progressive Taxation *H. Peyton Young* 253
Job Accessibility and Racial Differences in Youth Employment Rates
Keith R. Ihlanfeldt and David L. Sjoquist 267
Sex Discrimination in Labor Markets: The Role of Statistical Evidence:
Comment *Debra A. Barbezat and James W. Hughes* 277
Comment *William E. Even* 287
Reply *Peter Kuhn* 290
Self-Selection and the Earnings of Immigrants:
Comment *Guillermina Jasso and Mark R. Rosenzweig* 298
Reply *George J. Borjas* 305
The Effects of Fiscal Policies When Incomes Are Uncertain:
A Contradiction to Ricardian Equivalence:
Comment *William A. Bomberger* 309

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. C. 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit scientific organization, located at 1313 21st Avenue South, Suite 809, Nashville, Davidson County, TN 37212–2786. The Editor is Professor Orley Ashenfelter, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542–4607. During the preceding 12 months the average number of copies printed for each issue was 27,900; the average paid circulation, 25,295; the average free distribution, 95; the average number of copies distributed, 25,390. Corresponding figures for the issue published nearest to the filing date: total number copies printed, 28,000; total paid circulation, 26,322; total free distribution 94; total distribution 26,416.

P 5779

- Submit manuscripts (4 copies), 50 pages maximum, single-sided, double-spaced, to:

Orley Ashenfelter, Editor, *AER*, 209 Nassau Street, Princeton, NJ 08542–4607.

- Submission fee: $50 for members; $100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

Editorial Statement

It is the policy of the *American Economic Review* to publish papers only where the data used in the analysis are clearly and precisely documented, are readily available to any researcher for purposes of replication, and where details of the computations sufficient to permit replication are provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

The Future of the Income Tax[†]

By JOSEPH A. PECHMAN*

The federal income tax has been under attack by the economics profession for more than a decade. The attack comes from two directions: supply-siders who believe that progressive income taxation impairs economic incentives,[1] and more traditional economists who would substitute a progressive expenditure tax for the income tax.[2] At one time, support for the expenditure tax was confined to a few members of our profession, including such distinguished names as John Stuart Mill, Irving Fisher, Nicholas Kaldor, and James Meade. Today, it is fair to say that many, if not most, economists favor the expenditure tax or a flat rate income tax. This group has joined the opponents of progressive taxation in the attack on the income tax.

Despite an incessant barrage from both groups, no country in the world is planning to abandon the income tax or is even considering a personal expenditure tax. A wave of tax reform, beginning with the U.S. reform in 1986, has been sweeping the world, aimed at improving the income tax, not at eliminating it. Tax preferences formerly regarded as sacrosanct are being removed and there is a distinct movement toward comprehensive income taxation.[3] However, individual income tax rates are being cut, tax progressivity has been declining almost everywhere, and reliance on the income tax has been diminishing.

It will come as no surprise to this audience that I approve of the base-broadening feature of the current tax reform movement, but I believe that the reduction in the redistributive effect of the income tax has gone too far. In this paper, I shall show that the progressivity of the U.S. tax system—never very pronounced, except during and immediately after the two world wars—has been declining for more than two decades and that the Tax Reform Act of 1986 reversed this decline, but only slightly. Consequently, we have a long way to go to improve the equity of the tax system. I believe this can be done without punitive tax rates that will hurt economic incentives.

I begin with a brief review of recent changes in the U.S. distribution of income and follow this with an analysis of the effect of taxes on the income distribution. I next examine arguments for and against the income tax, with particular emphasis on its effects on economic incentives and its merits when compared with the expenditure tax. I then evaluate the income tax as it emerged from the 1986 tax reform and conclude with an agenda for further reform in the context of the current fiscal crisis. I believe that, when the nation gets around to eliminating or substantially reducing the federal deficit, the income tax should play an important role.[4]

[†] Joseph Pechman passed away on August 19, 1989. His Presidential address was delivered at the one-hundred second meeting of the American Economic Association, December 29, 1989, Atlanta, Georgia.

*Joseph A. Pechman, Economic Studies Program, The Brookings Institution, 1775 Massachusetts Avenue NW, Washington, DC 20036. I have benefited from the comments and suggestions on an earlier draft of this paper by Henry J. Aaron, Richard Goode, Jane G. Gravelle, Robert W. Hartman, Donald W. Kiefer, Herbert E. Klarman, Robert D. Reischauer, Clifford M. Winston, and H. Peyton Young, but they should not be held responsible for the views expressed in this paper. I am indebted to Richard Kasten for various simulations. I am also grateful to Stephen J. Kastenberg for research assistance and to Diane A. Shugart, Valerie M. Owens, and Sara C. Hufham for secretarial assistance.

[1] Some of the more extreme supply siders argued that large tax cuts pay for themselves (see, for example, Laffer, 1981), but I believe it is fair to say that this view has been totally discredited. For a more reasonable supply-side view, see the *Economic Report of the President 1982*.

[2] See, for example, Michael J. Boskin (1978), David E. Bradford (1980), Charles L. Ballard, Don Fullerton, John Shoven and John Whalley (1985), Paul Courant and Edward Gramlich (1982), Martin Feldstein (1978), Robert Hall and Alvin Rabushka (1985), John Kay and Mervyn King (1983), Charles McLure (1987), Mieszkowski (1980), and Lawrence Summers (1981). It is interesting that the recent popularity of the expenditure tax among economists was stimulated by a tax lawyer, William D. Andrews (1974).

[3] See Pechman (1988).

Distribution of Income and Tax Burdens

It is well known that, after several decades of relative stability, the U.S. pre-tax income distribution has become much more unequal in the last ten years. Official statistics understate the increasing inequality. At the same time, the tax system as a whole—and the income tax in particular—has become less equalizing, so that the trend toward inequality is even more pronounced after tax than before tax.

Distribution of Income. The longest continuous and comparable income distribution series available to us comes from the annual Current Population Survey (CPS) of the Census Bureau. The figures show that the share of total income received by the highest fifth of the nation's families fell from 1948 to 1952, remained unchanged between 1952 and 1981, and then rose from 1981 to 1988. By 1988, the share of the top fifth was the highest ever recorded. The figures for the top 5 percent are similar, except that their share in 1987 had not quite recovered to the 1952 high (Table 1).

It is well known that very high incomes are virtually unrepresented in the CPS distribution and that official census statistics greatly understate income inequality in any year. What is not recognized is that the CPS data greatly understate the *increase* in inequality that has occurred during the 1980s because very high incomes have been increasing much faster than the incomes in the lower part of the distribution.[5] This can be seen by examining changes in the shares of the top income recipients reported in the annual *Statistics of Income* published by the Internal Revenue Service (Table 2).[6]

TABLE 1—BEFORE-TAX INCOME SHARES, CENSUS DATA, SELECTED YEARS, 1948–1988, PERCENT.

Year	Top 5 percent of families	Top 20 percent of families
1948	17.1	42.4
1952	17.4	41.5
1957	15.6	40.4
1962	15.7	41.3
1967	15.2	40.4
1972	15.9	41.4
1977	15.7	41.5
1981	15.4	41.9
1987	16.9	43.7
1988	17.2	44.0

Source: Bureau of the Census. Income includes transfer payments (for example, Social Security benefits, unemployment compensation, welfare payments, etc.) but excludes capital gains. Distribution includes only families and excludes single persons living alone.

Like the CPS data, the tax data show that the very rich in the United States—defined as either the top 1 percent or the top 5 percent of the income distribution—enjoyed about the same income increases as the average income recipient in the 1950s, 1960s, and 1970s, but their share of total income has been rising in the 1980s. From 1952 to 1981, the share of the top 1 percent of the tax units remained in a very narrow range—between 8 and 9 percent of the total income reported on tax returns. Since 1981, their share has skyrocketed to 14.7 percent in 1986. The same trends are shown by the top 2, 5, 10, and 15 percent of the tax units.

Much of the increase in the share of the top tax units reflects the large increase in realized capital gains that accompanied the bull market of the 1980s. But salaries and other incomes of the top units have also been increasing faster than average.[7] In fact,

---

[4] See Musgrave (1989) for a statement of similar views.

[5] Using Pareto distributions based on income tax data to approximate the upper tail of the U.S. distribution, Rudy Fichtenbaum and Hushang Shahidi (1988) calculated that the CPS underestimation of the Gini coefficient rose from 1.7 percent in 1967 to 7.6 percent in 1984.

[6] For the method of calculation, see Pechman (1989), ch. 1.

[7] According to *Statistics of Income*, the shares of adjusted gross income other than capital gains increased from 6.59 percent in 1981 to 7.64 percent in 1986 for

TABLE 2—BEFORE-TAX INCOME SHARES, TAX DATA, SELECTED YEARS, 1948–1986, PERCENT

Year	Top 1 percent of Tax Units	Top 2 percent of Tax Units	Top 5 percent of Tax Units	Top 10 percent of Tax Units	Top 15 percent of Tax Units
1948	9.8	13.4	20.2	27.9	34.3
1952	8.7	12.1	18.7	26.7	33.4
1963	8.8	12.3	19.4	28.2	35.5
1967	8.8	12.3	19.6	28.3	35.5
1972	8.0	11.4	18.7	27.8	35.4
1977	7.8	11.3	18.9	28.3	36.1
1981	8.1	11.5	19.0	28.6	36.5
1986	14.7	18.2	26.6	36.8	45.1

Source: Statistics of Income. Income excludes transfer payments, but includes realized capital gains in full.

the movement toward inequality must have been even greater than the tax data show because they do not include the large amounts of income taxpayers were able to shelter before the enactment of the Tax Reform Act of 1986.

Many economists and statisticians have examined these trends, but nobody has been able to explain them fully. The declining share of incomes received by the lower income classes has been attributed to the increase in the number of single-parent families, slow growth in earnings of production workers, the disappearance of middle-income jobs, and other factors.[8] But these explanations do not account for the recent explosion of earned and property incomes of those in the top tail of the distribution.

The trend toward greater inequality has developed despite the existence of an income tax in the United States for seventy-six years and of an estate tax for eighty years. Clearly, the tax system never reduced inequality very much and other forces in the 1980s have swamped whatever equalizing effect it may have had earlier. I turn now to an examination of the burdens imposed by the tax system and how they have affected the distribution of income after tax.

Distribution of Tax Burdens. I have been estimating federal, state, and local tax burdens by income classes for the last two decades on the basis of the Bookings MERGE files.[9] These files are based on the CPS surveys, modified at the top by the incomes reported on federal individual income tax returns. As shown in Table 3, the tax burdens of the bottom 90 percent of the income distribution did not change very much from 1966 to 1985. By contrast, the tax burdens of the top ten percent of income recipients fell, especially those of the top 5 percent and 1 percent. Effective tax rates of the top 5 percent dropped by one-fifth between 1966 and 1985 (from 32.7 percent to 26.0 percent); for the top 1 percent, the reduction was more than one-third (from 39.6 percent to 25.3 percent).

Tax burdens of the highest income recipients fell because top federal individual tax rates were reduced throughout this period, from 70 percent in 1966 to 50 percent in 1985. Furthermore, the federal corporation income tax dwindled to relative obscurity, falling from 4.1 percent of GNP in 1966 to 1.6 percent in 1985. The proliferation of personal deductions (for example, state and local taxes, interest payments, and IRAs), tax-exempt bonds, and tax shelters were also major factors in the reduction of the tax burdens in the top part of the income distribution. The reduction in the corporate tax

---

the top one percent of tax units, from 17.31 to 18.77 percent for the top 5 percent, and from 35.11 to 36.56 percent for the top 15 percent.

[8] See, for example, Frank Levy (1988) and Sheldon Danziger, Peter Gottschalk, and Eugene Smolensky (1989).

[9] See Pechman and Benjamin Okner (1974) and Pechman (1985).

TABLE 3—EFFECTIVE RATES OF FEDERAL, STATE, AND LOCAL TAXES, BY POPULATION PERCENTILES, SELECTED YEARS, 1966–1988[a]

	Percent					
Population Percentile[b]	1966	1970	1975	1980	1985	1988[c] (est.)
1st Decile[d]	16.8	18.8	19.7	17.1	17.0	16.4
2nd Decile	18.9	19.5	17.6	17.1	15.9	15.8
3rd Decile	21.7	20.8	18.9	18.9	18.1	18.0
4th Decile	22.6	23.2	21.7	20.8	21.2	21.5
5th Decile	22.8	24.0	23.5	22.7	23.4	23.9
6th Decile	22.7	24.1	23.9	23.4	23.8	24.3
7th Decile	22.7	24.3	24.2	24.4	24.7	25.2
8th Decile	23.1	24.6	24.7	25.5	25.4	25.6
9th Decile	23.3	25.0	25.4	26.5	26.2	26.8
10th Decile	30.1	30.7	27.8	28.5	26.4	27.7
Top 5 Percent	32.7	33.0	28.4	28.9	26.0	27.4
Top 1 Percent	39.6	39.0	29.0	28.4	25.3	26.8
All Deciles[e]	25.2	26.1	25.0	25.3	24.5	25.4

Source: Brookings MERGE files.

[a] Assumes corporate income and property taxes are borne by capital income.

[b] Arrayed by comprehensive income which includes transfer payments, employee fringe benefits, net imputed rent, and corporate earnings allocated to shareholders.

[c] Projected from 1985 on the basis of CBO estimates of changes in effective federal tax rates. Assumes no change in effective state-local tax rates between 1985 and 1988.

[d] Includes only units in the sixth to tenth percentiles.

[e] Includes negative incomes not shown separately.

reflected primarily the investment incentives introduced in the 1960s and liberalized in the 1970s and 1980s, as well as a reduction in the profitability of the corporate sector.[10]

Since 1985, the distribution of tax burdens has changed largely because of the enactment of the landmark Tax Reform Act of 1986. This act increased the progressivity of the tax system, most notably by raising the personal exemptions, standard deductions, and the earned income credit, and by shifting about $25 billion of tax annually from individuals to corporations. However, this change in tax policy restored only a small fraction of the progressivity lost in the preceding two decades. At the very top of the income distribution, the 1986 federal tax reform restored about half the reduction in effective tax rates between 1980 and 1985, but left them far below the 1966 levels: the top 1 percent paid only 26.8 percent in taxes in 1988 as compared with 39 percent in 1970 (Table 3).

The inescapable conclusion from these figures is that the well-to-do in our society had very large reductions in tax rates in recent years, while the tax rates at the low and middle income levels have not changed much. Since the before-tax distribution has become much more unequal in the 1980s, it follows that inequality has increased even more on an after-tax basis.[11]

Transfer Payments. The other major element of government policy affecting the distribution of income is the system of transfer payments, or negative taxes. This system in-

---

[10] See Alan Auerbach and James Poterba (1987).

[11] Since the income tax affects economic behavior, the before-tax distributions reflect the responses of taxpayers, through portfolio shifts, substitution of nontaxable fringe benefits for cash compensation, and other actions, to the level and structure of the tax system. Hence, the after-tax distribution is more representative of the changes in inequality. For a discussion of these issues, see Eugene Steuerle (1985) and Harvey Galper, Robert Luce, and Eric Toder (1988).

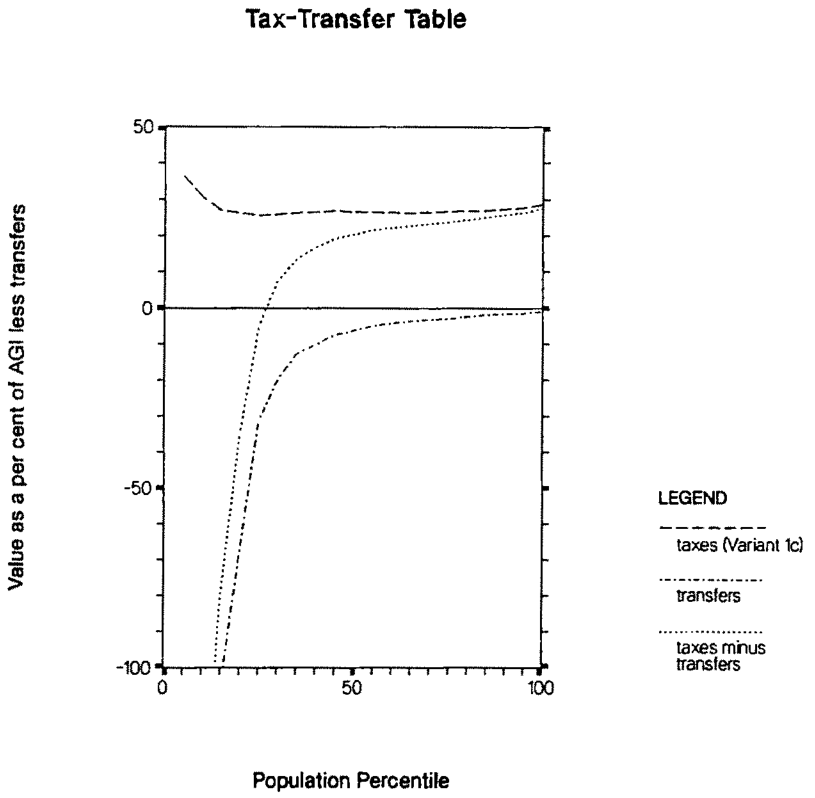


FIGURE 1. FEDERAL, STATE, AND LOCAL TRANSFERS AND TAXES AS A PERCENT OF MARKET INCOME BY INCOME PERCENTILE, 1985

cludes programs of public assistance that are designed explicitly to help the poor, but it also includes others that are not designed primarily for this purpose (for example, retirement and unemployment benefits and health insurance). To evaluate the impact of the tax-transfer system on the distribution of income, cash and in-kind transfers must be added to market incomes while taxes are deducted.

While I cannot separate the effects of transfer payments and of taxes on the recent changes in the after-tax income distribution,[12] a snapshot for a recent year—1985—suggests what happened (Figure 1). When family units are arrayed by their incomes from market production (wages, salaries, interest, dividends, etc.), the U.S. tax system is only mildly progressive. On the other hand, transfer payments are highly progressive. Taxes in 1985 were regressive in the lowest deciles and proportional thereafter, while transfer payments declined from over 200 percent of market incomes in the lowest decile to 1.4 percent in the highest. On balance, families in the lowest three deciles received more in transfers than they paid in taxes, while those in the top seven deciles paid more in taxes than they received in transfers.

Clearly, the tax-transfer system is progressive, mainly because of transfers, not taxes. What we are doing in the United States is financing redistributive transfers with taxes that are roughly proportional to incomes. Moreover, the tax system has been getting less progressive in the last two decades, while the ratio of transfers to income has been increasing.[13] In other words, the recent increases in transfer payments in the United States have been financed by the low and middle income groups, while the rich have been getting tax cuts.

---

[12] The income distributions used in the Brookings MERGE file for 1966 were prepared on the basis of incomes *including* transfer payments. After so many years, it is impractical to reclassify the income and tax data for 1966 by income excluding transfer payments.

What Role for the Income Tax?

Most people support tax progressivity on the ground that taxes should be levied in accordance with ability to pay, which is assumed to rise more than proportionately with income. Economists have long had trouble with the "ability to pay" concept. In recent years they have revived the old notion that consumption measures ability to pay better than income does. I believe that the person in the street is right and that we should continue to rely on the income tax to raise revenue in an equitable manner.

Ability to Pay. In the latter half of the nineteenth century, progressive income taxation was justified by "sacrifice" theories that emerged from discussions of ability to pay. Under this doctrine, ability to pay is assumed to increase as incomes rise, and the objective is to impose taxes on a basis that would involve "equal sacrifice" in some sense. If the marginal utility of income declines more rapidly than income increases and the relation between income and utility is the same for all taxpayers, equal sacrifice leads to progression.[14] Whether or not one believes in sacrifice theory, the ability to pay idea has been a powerful force in history and has doubtedly contributed to the widespread acceptance of progressive taxation.[15] Young has found that the equal sacrifice model fits most U.S. tax schedules in the postwar period, with the notable exception of the schedule adopted in 1986. Similar results hold for Italy, West Germany, and Japan.[16]

Henry Simons vigorously attacked sacrifice theory although he argued strongly that the purpose of the progressive income tax is to reduce economic inequality.[17] Simons was vague on how far progression should be pushed, but he clearly felt that it had not yet gone too far in most countries. His prescription was the pragmatic one that the tax rates should not impair economic incentives. In his policy statements, he argued in favor of a broad base and graduated rate schedule that rises to a maximum of 50 percent.[18]

I agree with Simons that the income tax should be used to reduce the great disparities of welfare, opportunity, and economic power arising from the unequal distribution of income. I also recognize that this view is not widely held and has probably not been the major rationale for income tax legislation in the United States or in most other countries. The income tax is widely used primarily because it raises large amounts of revenue in a moderately progressive way. Recent income tax reforms have concentrated mainly on eliminating tax preferences to improve horizontal equity; where income tax rates had been pushed to very high levels, they are being moderated. Curiously, the world appears to be moving toward a consensus on the Simons' view that the income tax should be levied on a broad base with graduated rates reaching a maximum of 50 percent or less, though not for his reasons.

---

[13] Little empirical work has been done on trends in the progressivity of the transfer system itself. If it has changed, the movement in the last decade may have been toward less progressivity because welfare payments have not kept up with inflation and the coverage of the unemployment compensation system has narrowed.

[14] To be precise, equal *absolute* sacrifice leads to progression, equal *proportionate* sacrifice to still more progression, and equal *marginal* sacrifice to leveling of incomes from the top down until the required revenues are obtained. See Richard Musgrave (1959), pp. 99–102.

[15] See Musgrave (1959), ch. 5 and Richard Goode (1976), ch. 2. For a skeptical view of the case for the progressive income tax, see Blum and Kalven (1952).

[16] See H. P. Young (1990). I omit discussion of the optimal tax literature, which built on the old sacrifice theory, because it does not yet provide a basis for making judgments about the optimum degree of progression. This literature suggests that a progressive income tax is appropriate when redistribution is an objective of social policy, but the range of tax rates is very wide depending on the assumptions used. See Glen Hubbard and Kenneth Judd (1986), Nicholas Stern (1987), pp. 49–52, Henry Aaron (1989), pp. 10–12, Rosen (1988) and Slemrod (forthcoming).

[17] "The case for drastic progression in taxation must be rested on the case against inequality—on the ethical or aesthetic judgment that the prevailing distribution of wealth and income reveals a degree (and/or kind) of inequality which is distinctly evil or unlovely." Simons (1938), pp. 18–19.

[18] See Simons (1950).

Economic Incentives. The effects of the progressive income tax on incentives to work and to save are hard to measure. As is well known, the substitution and income effects of taxation work against each other, and the net result cannot be predicted.

Sample surveys have revealed that professional personnel do not vary their hours of work in response to high tax rates.[19] However, recent econometric studies suggest that the pre-1986 income and payroll taxes reduced the work effort of primary earners in the United States by about 8 percent, while secondary earners—who have a greater opportunity to vary their labor input—reduced their work effort by as much as 30 percent.[20] According to this approach, the 1986 reform may have increased labor supply of married men by only 1 percent and of married women by less than 3 percent, largely because the marginal tax rates of most workers were not reduced very much.[21] Burtless estimates that the Reagan tax and transfer policies increased average annual taxes of men aged 25–54 by no more than 2–4 percent and of women in the same age group by no more than 3.5 percent.[22]

Historical trends in U.S. labor supply are not consistent with the finding that taxes have reduced work effort. Adult males have been reducing their labor supply over the last forty years, largely through earlier retirement little of which is the effect of tax rates. The labor force participation of women has risen sharply in recent years, despite high marginal rates resulting from the requirement that married couples must file joint returns to benefit from income splitting. Studies in other countries are not reliable enough to support conclusions about the relationship between taxes and labor supply.

The effect of taxes on saving is even more ambiguous. A few studies claim that they have found a significant response to an increase in the real after-tax return on saving; others find that the response, if any, is close to zero.[23] The reduction in the personal saving rate in the United States in the 1980s confounded most economists in view of the reductions in the marginal tax rates, the incentive provided by individual retirement accounts (IRAs), and the high real interest rates, all of which should have increased the incentive to save.

The strongest conclusion one can draw from the available evidence is that the incentive effects of taxation have been relatively small. Yet the supply siders were convinced that the incentive effects were so large that rate cuts would increase revenues when tax rates are reduced.[24] U.S. tax rates were cut sharply in 1981 and 1986, but these cuts had little effect on labor supply and no effect on saving. Under the circumstances, so long as tax rates are not pushed to punitive levels, incentive considerations do not justify neglect of the distributional objective of tax policy.

Income vs. Expenditure Tax. The revival of interest in the expenditure tax can be traced to the difficulties of taxing income from capital under the income tax. However, economists and tax lawyers have also found efficiency reasons to prefer the expenditure tax and these need to be addressed.[25]

---

[19] See Break (1957) and Holland (1969).

[20] These estimates are based on a comparison with work effort under a system of lump-sum taxes. The estimates would be reduced substantially if the basis of comparison were a proportional income tax. For summaries of recent studies and their implications for policy, see Bosworth (1984), ch. 5 and Gary Burtless and Robert Haveman (1987). The methodology of these studies is explained by Hausman (1987).

[21] See Hausman and Poterba (1987).

[22] See Burtless (1989).

[23] For a review of the various studies, see Evans (1983) and Bosworth (1984), chaps. 3 and 5.

[24] President Ronald Reagan was one of those who believed that tax rate cuts would actually increase revenues (see Reagan, 1981, p. 710). The supply-side view is defended in Raboy (1982). For an analysis of this view, see Meyer (1981).

A basic difference between the income and expenditure taxes is in the time perspective of the two taxes. The perspective of the income tax is relatively short run—a year or several years to allow for short-run income fluctuations. Consumption is more stable than income and is alleged, therefore, to be a better measure of long-term well-being. In fact, under certain simplifying assumptions, the *bases* of taxes on the discounted present value of income and expenditure are the same over a lifetime. Assuming perfect capital markets, constant discount rates that apply equally to all people under all circumstances, tax rates that are constant and proportional, and no gifts and bequests, the present values of lifetime expenditures of people with the same (discounted) lifetime incomes are the same regardless of when the incomes are consumed.[26]

Advocates of the expenditure tax regard the lifetime perspective as a major advantage because it permits them to pretend that taxing consumption is equivalent to taxing personal endowments. A tax on endowments, if they could be measured, would avoid the distortionary effects of either an income tax or an expenditure tax. If it is assumed that lifetime consumption approximates endowment, then taxing consumption at flat and constant rates treats equally all taxpayers with the same endowment.[27] The totally unrealistic assumptions underlying this line of reasoning strain credulity, but it does seem to lie behind the strong support for the expenditure tax by many economists.

The lifetime perspective has little merit even without the endowment rationale. In my view, it is difficult enough to measure economic circumstances over relatively short periods. Taxation of lifetime consumption (or income) hardly seems appropriate in a world of changing tax rates, substantial family instability, economic and political change, and uncertainty. Except for the attractiveness of the arithmetic, lifetime economic circumstances as measured by discounted lifetime incomes or consumption cannot be regarded as satisfactory indexes of ability to pay.[28] Moreover, taxation of annual consumption expenditures at graduated rates would destroy the identity of lifetime taxes of taxpayers with the same (discounted) lifetime incomes.

The expenditure tax is alleged to be superior to the income tax on the additional ground that the income tax reduces the return on saving and therefore encourages current as against future consumption. Even if saving remained unchanged, the distortion generates a welfare loss for consumers. It has been pointed out by many economists that this effect must be balanced against the welfare cost of further distorting the choice between labor and leisure. There is no theoretical basis for judging whether the welfare gain from eliminating the intertemporal distortion of consumption would exceed the welfare loss from increasing the intratemporal distortion of the labor-leisure choice.[29]

A tax that omits saving from the tax base can be shown to be the same as a tax applying only to labor income and exempting all property income.[30] Several expenditure tax advocates have, in fact, proposed a tax on labor income on grounds of simplicity and administrative feasibility.[31] Most people would be appalled by a proposal to substitute a wage tax for income tax, yet that is essentially what expenditure tax proponents are advocating.

---

[25] For analyses of the relative merits of the income and expenditure taxes, see Kaldor (1955), Musgrave (1976), Pechman (1980), Aaron and Galper (1985), and Bradford (1986).

[26] The identity of lifetime incomes and expenditures holds when there are gifts to others and bequests, provided the gifts and bequests are counted as expenditures. Aaron and Galper (1985) defined such a tax as a "lifetime" income tax, arguing that "there is no logical reason why a particular astronomical regularity should be enshrined in the tax law" (pp. 21–22).

[27] See Bradford (1980), pp. 106–9, Bradford (1986), p. 315, and Musgrave (1976), pp. 11–12. Bradford holds this view while Musgrave does not.

[28] For an excellent discussion of the lifetime vs. the short-run perspective as a basis for taxation, see Goode (1980).

[29] See, for example, Feldstein (1978), Atkinson and Sandmo (1980), Atkinson and Stiglitz (1980), Auerbach and Kotlikoff (1987), ch. 5, Rosen (1988), p. 469, and Gravelle (1988 and 1989).

[30] Like the equality between a consumption and an income tax in present value terms, this assumes that tax and interest rates remain constant over a lifetime and there are no gifts and bequests. Let $s =$ saving, $w =$ wages, $c =$ consumption, $r =$ the interest rate, and assume that saving in the first period is consumed during

Many economists are attracted to the expenditure tax because it would not tax income from capital and would thus eliminate all the income tax problems arising from the use of the realization principle for calculating capital gains and losses and from the accounting conventions for inventories, depreciation, and depletion used in arriving at net business profits. There would also be no need to adjust the tax base for inflation, as consumption would be measured appropriately in current dollars. These are serious problems for income taxation and I shall deal with them later, but it would be unfortunate to abandon the income tax for administrative and compliance reasons alone.

The transition from the income tax to an expenditure tax would be troublesome. The retired elderly would draw down assets, some of which had previously been taxed under the income tax, to finance current consumption that would be taxed yet again. To avoid this double tax, some method would need to be devised to identify consumption from previously taxed accumulations. Grandfathering all assets at the time an expenditure tax is initiated would leave a big loophole for people with large amounts of untaxed accrued capital gains. But I have not seen any practical method of making the necessary distinctions in order to prevent wholesale tax avoidance and to achieve equity.[32]

Under an expenditure tax, taxpayers who save could accumulate large amounts of wealth over a lifetime. Many, but by no means all, expenditure tax advocates support wealth or estate and gift taxes to prevent excessive concentrations of wealth. But the history of transfer taxation in this country and abroad provides little assurance that effective death and gift taxes would be levied to supplement an expenditure tax.

Proponents of expenditure taxation often compare the merits of a comprehensive expenditure tax with the income tax as it has developed. It is hard to believe that an expenditure tax would be enacted without numerous exemptions and exclusions. In fact, most of the eroding features of the income tax (for example, preferences for housing, fringe benefits, child care, state-local borrowing, etc.) might be carried over to the expenditure tax. Thus, an expenditure tax is no less immune to erosion than the income tax and, in such circumstances, it loses much of its attractiveness.

I conclude that income is a better indicator of ability to pay than consumption and that the major upheaval of substituting an expenditure tax for an income tax cannot be justified on theoretical or practical grounds.

How Much Progression? The effective degree of progression of the income tax depends on the comprehensiveness of the tax base as well as on the tax rates. We have learned from experience that high, graduated tax rates do not assure progressivity of the income tax. For most of the period since the end of World War II, the top U.S. income tax rates were 70 percent or higher. Yet little equalization resulted because of the erosion of the base of the individual and corporate income taxes and because of increases in the payroll tax for Social Security.[33] According

---

the second, then

$$s_1 = w_1 - c_1$$

$$c_2 = w_2 + (1+r)s_1$$

$$c_1(1+r) + c_2 = (1+r)(w_1 - s_1) + w_2 + (1+r)s_1$$

$$c_1(1+r) + c_2 = (1+r)w_1 + w_2.$$

[31] See Hall and Rabushka (1985), Bradford (1986), ch. 14, and McLure, Mutti, Thuroni, and Zodrow (1988), ch. 9.

[32] For a discussion of these problems and possible solutions, see Aaron and Galper (1985), pp. 78–79 and 99–103, and U.S. Department of the Treasury (1977), chaps. 4 and 6.

[33] The combined employer-employee payroll tax rose from 6 percent in 1960 to 12.26 percent in 1985 and

TABLE 4—GINI COEFFICIENTS OF INEQUALITY BEFORE AND AFTER FEDERAL TAXES, SELECTED YEARS, 1977–1988

	Gini Coefficients	
Year	Before Federal Taxes	After Federal Taxes
1977	0.4502	0.4185
1980	0.4627	0.4320
1984	0.4884	0.4700
1988	0.4940	0.4722

Source: Congressional Budget Office, *The Changing Distribution of Federal Taxes: A Closer Look at 1980*, (July 1988), p. 98.

to estimates of the Congressional Budget Office (CBO), in 1977, when the top income tax rate was 70 percent and the general corporate tax rate was 48 percent, the Gini coefficient of inequality was 0.4502 before tax and 0.4185 after federal taxes (more than half of which were individual and corporate income taxes). In 1980, when the top tax rate was still 70 percent (though only on unearned income) and the corporate rate was 46 percent, the Gini coefficient was 0.4627 before tax and 0.4320 after the same taxes. Thus, as measured by the Gini coefficient, the equalization achieved by the federal tax system declined from a modest 0.0317 points (7.0 percent) in 1977 to an even more modest 0.0307 points (6.6 percent) in 1980 (Table 4).

Two major pieces of tax legislation were enacted during the 1980s. One increased inequality, the other reduced it. The 1981 Act, enacted after Ronald Reagan's sweeping victory in the 1980 presidential election, increased inequality by reducing income tax rates by 23 percent across the board (with a top rate on ordinary income of 50 percent), lowering the capital gains rate to a maximum of 20 percent, introducing generous deductions for individual retirement accounts, and providing very liberal depreciation allowances for business investment on top of the previously enacted investment tax credit. The Tax Reform Act of 1986 reduced inequality by increasing personal exemptions and the standard deduction, equalizing the tax rates on capital gains and ordinary income, and closing numerous loopholes. At the same time, income tax rates were reduced to a maximum of 33 percent on individuals and 34 percent on corporations.

Despite the large rate cuts at the top of the income scale, the 1986 act increased income tax progression, though not to the 1980 level. By 1984, the equalization provided by the federal tax system had declined to 0.0184 points (3.8 percent) in terms of the Gini coefficient. As a result of the 1986 act, the degree of equalization increased to 0.0218 points (4.4 percent) in 1988 (Table 4). While this change is modest, it is noteworthy as the first movement toward greater income tax progressivity at least since 1964, when the Kennedy-Johnson tax cut was enacted.

I suggest that a minimal goal of federal tax policy in the next several years should be to restore the equalization achieved by the federal tax system in the mid-1970s.[34] While this may appear to be a modest goal, it turns out to be a rather ambitious undertaking, particularly if income tax rates are to be kept to moderate levels. Before calculating the tax rates, it is necessary first to establish the appropriate tax base for a modern income tax.

Reform of the Income Tax

The proper base for the income tax was described fifty years ago by Henry Simons, who argued that it should conform with an economic definition of income.[35] Admittedly, the use of a comprehensive income tax con-

---

14.1 percent in 1985. It is scheduled to rise further in 1990 to 15.3 percent. The maximum taxable earnings level rose from $4,800 in 1960 to $7,800 in 1970, $25,900 in 1980, $39,600 in 1985, and $48,000 in 1989. The maximum taxable earnings levels have been adjusted for inflation since 1984. For details see Pechman (1987), p. 332.

[34] Specifically, as the basis for revision, I use the distribution of federal tax burdens in 1977 when the CBO tax incidence series begins. I do not mean to imply that the federal tax system in the mid-1970s had exactly the correct degree of equalization.

[35] See Simons (1938), ch. II.

tradicts the principle of optimal taxation that tax rates should vary with a number of elasticities. However, the optimal tax models are based on strong assumptions that are often implausible or virtually impossible to validate. Consequently, there is no empirical basis for determining how different commodities and sources of income should be taxed. Moreover, the compliance and enforcement costs of such a system could be large enough to more than offset the potential inefficiencies of a uniform tax. In the absence of reliable data, it is safer to rely on the comprehensive approach rather than to introduce tax differentials that will generate their own distortions.[36]

According to Simons and others, income is the sum of an individual's consumption and change in net worth during a particular time period. For a long time, the federal income tax base was a far cry from a comprehensive definition of income. In 1986, however, Congress reversed its previous practice and enacted a wholesale tax reform that moved the income tax a long way toward the Simons' ideal. This remarkable piece of legislation can provide the basis for achieving the distributive objectives discussed earlier with moderate tax rates.

The 1986 Tax Reform.[37] The Tax Reform Act of 1986, a major step toward comprehensive income taxation, greatly improved the fairness and efficiency of the tax system. The major accomplishments of the act are as follows:

By doubling personal exemptions and increasing the standard deduction, the act relieved about 5 million poor people from paying any income tax. This step restored the principle (abandoned by Congress in 1978) that people who are officially designated as "poor" should not be required to pay income tax. The principle was perpetuated by the resumption in 1989 of an automatic annual adjustment of the exemptions and standard deduction for inflation.

Significant increases were made in the earned income credit for wage earners with families. These increases eliminated almost the entire Social Security payroll tax (including the employer's share) for those eligible for the full credit and reduced the tax burden for many low-income workers.

For the first time since 1921, realized capital gains were made taxable as ordinary income. This is the keystone of comprehensive tax reform: it reduces the incentive to convert ordinary income into capital gains and removes one of the major elements of tax shelter arrangements. Moreover, this change made it possible to reduce tax rates without reducing the progressivity of the income tax.

A good start was made to reverse the erosion of the individual income tax base. For example, unemployment benefits, which were previously taxable only if a married taxpayer's income exceeded $18,000 ($12,000 if single), were made taxable regardless of the size of income. Deductions for state and local sales taxes were eliminated and those for consumer interest were phased out. For administrative reasons, deductions for unreimbursed business expenses, costs incurred in earning investment income, and other miscellaneous expenses were allowed only to the extent that they exceed a floor of two percent of income.

Some of the most egregious loopholes and special tax benefits were eliminated. Many tax shelters were rendered unprofitable by denying deductions for losses from passive activity against income from anything but passive activities.[38] Tax subsidies for borrowing (other than for mortgages) were eliminated by another limitation on the deduction for interest expenses to the amount of investment income reported on the individual's tax return.[39] Deductions for contributions to individual retirement accounts were curtailed. Deductible business expense accounts for meals, travel and entertainment were limited to 80 percent of outlays. Tax preferences benefiting defense contractors, banks, oil companies and other industries were narrowed. On top of all these changes, the minimum tax for both individuals and business was retained and strengthened.

---

[36] See Stern (1987), p. 51, Aaron (1989), pp. 10–12, and Slemrod (forthcoming).

[37] For an analysis of the structural features of the 1986 tax reform, see Pechman (1987a).

[38] A passive activity is a trade or business in which the taxpayer (or spouse) does not materially participate. All rental activities are regarded as passive.

[39] However, the act did not change the deductibility of interest on borrowing to finance business investments.

Finally, the individual and corporate tax rates were cut drastically. Under the individual income tax, two rates—15 and 28 percent—were substituted for the earlier schedule of 14 rates, which rose to a maximum of 50 percent. However, the benefits of the lowest rates and of the personal exemptions were phased out for higher income taxpayers at a 5 percent rate. As a result, the new individual income tax rate structure has four brackets, with rates of 15, 28, 33, and 28 percent (see Table 7). The general corporate rate was cut from 46 percent to 34 percent. Despite these large rate cuts, the act was expected to be roughly revenue neutral in total over the first five years, but to shift about $25 billion of tax annually from individuals to corporations.

The distributional effect of the 1986 act is distinctly progressive, especially if the increase in corporate tax liabilities is taken into account. I have calculated the change in average effective tax rates of the nation's families on the basis of the distribution of income estimated from the Brookings MERGE file (Table 5). Total federal tax burdens decline in the lower nine deciles and rise in the top decile. In the lower deciles, the tax reductions result from increases in the personal exemptions, standard deduction, and the earned income credit under the individual income tax. The increases at the top reflect the broadened individual income tax base, as well as the increase in corporate tax liability, which is assumed to fall on owners of capital in these calculations. However, as already noted, this increase in progressivity only partially reversed the reductions that had taken place in the 1970s and early 1980s.

TABLE 5—CHANGES IN INDIVIDUAL AND CORPORATE INCOME TAX LIABILITIES UNDER THE TAX REFORM ACT OF 1986, BY POPULATION PERCENTILE, 1988

Population Percentile[a]	Percent Change in Federal Individual and Corporate Income Taxes[b]	Percent Change in Total Federal Taxes
1st decile	−44	−16
2nd decile	−32	−11
3rd decile	−24	−10
4th decile	−16	−7
5th decile	−12	−6
6th decile	−8	−4
7th decile	−7	−4
8th decile	−6	−3
9th decile	−6	−4
10th decile	+3	+2
Top 5 Percent	+4	+3
Top 1 Percent	+5	+5

Source: Brookings MERGE file.

[a] The classification is by a comprehensive definition of income, including imputed rent and corporate earnings allocated to stockholders, whether distributed or not.

[b] Assumes the corporate tax is on capital in general.

The Unfinished Agenda. Despite the progress made in 1986, the federal income tax in the United States falls considerably short of the comprehensive income target.[40] I assume that we shall continue to tax capital gains on a realization, rather than accrual, basis, and that gifts and bequests will be taxed under a separate transfer tax. Nevertheless, a great deal more could be done to broaden the tax base for equity, efficiency, and revenue reasons.

The personal exemptions, standard deductions, and rate bracket limits are adjusted annually for inflation, but the tax base is not. Of the two types of adjustment, adjustment of the tax base would be by far the more important. Perhaps the major reason why the income tax tends to be in disrepute is the discrimination against capital income inherent in a nominal income tax. An inflation adjustment of asset prices should be incorporated in the tax law as part of the computation of real capital gains and losses, real interest income expense, and real inventory and depreciation allowances. The adjustment of interest is admittedly difficult, but the widespread use of computers should ease the administrative and compliance problems.

---

[40] For a detailed analysis of a comprehensive income tax base, see Pechman (1977). Estimates of the revenue effects of specific changes in the tax base needed to reach the comprehensive base are given in Congressional Budget Office (1989).

Restoration of a tax differential between capital gains and ordinary income should be resisted at all costs. Equalization of the tax rates lowers the incentive to convert other income into capital gains, simplifies business and financial decisions, and reduces income tax complexity. Aside from the correction for inflation, the one additional reform needed in the capital gains tax is to include in the tax base unrealized capital gains transferred by gift or at death. Taxing such gains would reduce the lock-in effect of the tax on transfers of assets and eliminate a source of horizontal inequity.

A major neglected problem in most countries is the erosion of the tax base from the exclusion of employee fringe benefits. Trade unions, as well as employers, staunchly defend the continued exclusion of fringe benefit income, but in fact the largest subsidies go to the highest paid employees. Loopholes for union members and other workers are no more defensible than those for the rich. Taxation of fringe benefits would encourage their conversion into cash compensation, thus giving employees more control over the disposition of their income and the choice of the providers of their services. Australia and New Zealand have shown the way to reform in this area by taxing fringe benefits (other than contributions to pension plans) at the corporate tax rate. This method of handling a difficult, but urgent, problem is simple and effective.

Social Security benefits continue to receive favorable treatment, even though the elderly can no longer be regarded as a disadvantaged group. Under current law, the medical insurance subsidies they receive are not subject to tax, and less than half of retirement and disability benefits is taxable to married couples with income above $32,000 ($25,000 if single). The value of the medical insurance subsidies should be subject to tax in full[41] and retirement and disability benefits should be treated like private pensions without any income thresholds, which would mean that roughly 85 percent of the benefits would be currently taxable.

The treatment of owner-occupied housing remains unsatisfactory. I assume that the exclusion of imputed rent from the tax base and the deduction of mortgage interest by most homeowners are sacrosanct, but it is possible to limit the encouragement of borrowing without promoting rearrangements of debt for tax purposes. The solution is to broaden the limitation on deductions of investment interest to include all interest payments. That is, a deduction for all interest payments would be allowed, but limited to the amount of reported investment income. To accommodate the home-owner lobby, the limit could be raised to net investment plus an arbitrary amount, say, $10,000—enough to take care of the vast majority of home owners. The broader interest limitation would remove the discrimination against borrowing for other purposes and the incentive to substitute home equity loans for other types of borrowing.

Deductions other than for interest are still too generous. The Simons' definition of income includes all sources of income, without any deductions for the uses of that income. For equity reasons, it is appropriate to permit deductions for such unusual expenses as medical payments and casualty losses. I would retain the deduction for state income taxes to moderate interstate tax differentials.[42] However, the property tax is largely a benefit tax and therefore should not be deductible. Nor is it necessary to allow a deduction for the first dollar of charitable contributions on incentive grounds. Little or no charitable giving would be lost[43] and much revenue would be gained or reductions in tax rates would be possible if the federal deduction for property taxes were disallowed and the deduction for charitable contributions were restricted to amounts in excess of two percent of income. In addition, the tax-exemption for interest on newly issued state and local bonds should be removed.

---

[41] The tax would be imposed on the insurance value of hospital and Medicare subsidies rather than on the dollar benefits actually received. See Congressional Budget Office (1989).

[42] See Pechman (1987b), pp. 267–69, for illustrations of aggregate tax rates with various combinations of federal and state tax rates.

Although income tax compliance is better in the United States than in most other countries, roughly 15 percent of individual income is still unreported, according to IRS estimates.[44] Extension of withholding to interest and dividends would improve compliance. Congress enacted a withholding system for these income items in 1982, but repealed it the following year under pressure from the financial institutions. Since information returns are required for annual interest and dividend payments of $10 or more, the marginal costs of compliance with a withholding system would be small.

One of the major features of the 1986 tax law was to telescope the schedule of tax rates into two acknowledged and two concealed brackets, a bizarre four-bracket rate structure of 15, 28, 33, and 28 percent. The reduction in the number of brackets was a response to the flat tax proposals that were being promoted when the tax reform bill began its journey through Congress, while the unsightly bulge in the rate schedule was motivated by revenue considerations. It is not necessary to return to 14 brackets, but there is room for more rate graduation without the bulge.

In this connection, consideration needs to be given to improving the structure of estate and gift taxes to compensate for their low average rates. These taxes were almost gutted by increases in the exemptions and reductions in the tax rates when income tax rates were cut in 1981. Now that the top income tax rates are even lower, it is time to rely more heavily on the estate and gift taxes.

The reduction in the tax rates led to two additional changes in the 1986 act that I believe were unfortunate. Congress eliminated the deduction for two-earner couples and ended the privilege of averaging income for tax purposes.[45] Both provisions should be restored in the interest of horizontal equity.

Finally, contrary to the prevailing view among public finance experts, Congress clearly believes that a separate, unintegrated corporate tax is essential for effective income taxation. A separate tax prevents individuals from avoiding the income tax by accumulating earnings at the corporate level, although some might question whether corporations should be taxed at a higher rate than the top bracket individual rate. But in its present form, the corporate tax encourages debt financing. It is alleged to be a major cause of the recent upsurge in leveraged buyouts and mergers. The remedy is not to allow a deduction or credit for dividends received at the individual level. Rather, the deduction of interest by corporations should be denied while reducing their tax rate to the neighborhood of 15 percent to maintain the revenues now produced by the corporate tax.[46] The corporate tax would become a low-rate tax on net corporate income before distributions.

Tax Rates and Progressivity. The reforms I have suggested would greatly increase the income tax base and permit a realignment of the tax rates to achieve the distributional objectives described earlier. At calendar year 1990 levels, the tax base would increase from $2.4 trillion to $2.8 trillion (Table 6). The increase in the base leaves enough room to cut rates in the lowest taxable income brackets and still keep the top tax rates at reasonably modest levels.

---

[43] Clotfelter (1989) reports finds that equations estimated on pre-1986 data relating charitable giving to income, the price of giving, and other variables fail utterly to predict the response of charitable giving to the Tax Reform Act of 1986.

[44] Kenadjian (1988).

[45] For discussions of the function of a two-earner deduction and the need for income averaging, see Pechman (1987a), pp. 102–7 and 127–28.

[46] To prevent undue hardship to highly leveraged firms, the denial of the interest deduction might be phased in over a period of years.

TABLE 6—ADJUSTED GROSS INCOME AND TAXABLE INCOME UNDER THE TAX REFORM ACT OF 1986 AND UNDER A COMPREHENSIVE INCOME TAX

Item	1990 Billions of Dollars	
	Adjusted Gross Income	Taxable Income
Tax Reform Act of 1986	3545	2407
Plus:		
Personal Deductions[a]	0	68
Transfer Payments[b]	226	164
Fringe Benefits[c]	187	185
Two-earner Deduction[d]	−82	−81
Other[e]	43	42
Equals: Comprehensive Tax[f]	3919	2785

Source: Congressional Budget Office.

[a] Allows flat standard deduction of $4,000; interest deduction limited to investment income plus $10,000; tax deduction limited to income taxes; 10 percent floor on deductions for medical expenses and casualty losses; 2 percent floors on deductions for charitable contributions and miscellaneous expenses; and no standard deduction for the elderly and the blind.

[b] Includes 85 percent of Social Security retirement and disability benefits for all taxpayers, workers' compensation, and veterans' benefits; and 50 percent of the insurance value of hospital insurance benefits.

[c] Includes premiums paid by employers for health and life insurance and other fringe benefits; interest on life insurance policies; and IRAs of persons covered by employer pension plans.

[d] 20 percent of earnings of spouses with lower earnings up to a maximum of $70,000.

[e] Includes unrealized capital gains transferred by gift or at death. interest on newly issued state and local securities, and all preference items now subject to the minimum tax.

[f] An increase in the earned income credit under Plan II (see Table 7) does not affect adjusted gross income or taxable income.

In redesigning the rate structure, I suggest scrapping the multiple schedule system which was developed to reduce the tax advantage of married couples relative to single people under income splitting. It is simpler to use one set of rates and to rely on the personal exemptions to take into account differences in ability to pay of families of different size.[47] The restored deduction for two-earner couples (20 percent of the earnings of the spouse with the lower earnings up to $70,000) would help avoid a significant marriage penalty.

The same revenue and progressivity of present law could be generated by a tax schedule ranging from 7 percent on the first $5,000 of taxable income to 26 percent on taxable income in excess of $35,000, without the bulge in tax rates under current law (see Plan I, Table 7).[48] Thus, a wide margin exists for increasing progressivity at the top of the income scale, while keeping rates moderate. To restore the progressivity of the federal tax system to its 1977 level, the range of graduation would have to be expanded and the rate of graduation increased. A starting rate of 4 percent on the first $5,000 of taxable income rising to 48 percent on tax-

---

[47] To avoid the old community property problem, (see Pechman, 1987b, pp. 102–7) the brackets in the rate schedules for married couples filing separate returns would be one-half the size of the brackets for single people.

[48] Exemptions and the standard deduction would remain the same as in 1989, that is, exemptions would be $2,000 per capita and the standard deduction would be $3,100 for single persons, $4,550 for heads of households, and $5,200 for married couples, all adjusted for inflation.

TABLE 7—COMPARISON OF TAX RATES UNDER THE TAX REFORM ACT OF 1986 AND UNDER TWO COMPREHENSIVE INCOME TAXES, 1990

Tax Reform Act of 1986[a]		Comprehensive Income Tax[b]		
Taxable Income	Rate (percent)	Taxable Income	Rate (percent)[c]	
			Plan I	Plan II
\$0–\$32,400	15	\$0–\$5,000	7	4
32,400–78,400	28	5,000–10,000	10	8
78,400–208,560[d]	33	10,000–20,000	15	14
208,560 and over	28	20,000–35,000	18	14
		35,000–70,000	26	23
		70,000–100,000	26	24
		100,000–150,000	26	29
		150,000–250,000	26	33
		250,000 and over	26	48

[a] For a married couple with two dependents. Separate rate schedules apply to single persons and heads of household.

[b] Applies to all taxpayers, regardless of marital and family status.

[c] Plan I maintains present progressivity of the federal income tax at 1990 levels; Plan II restores progressivity of the federal tax system as a whole to levels prevailing in 1977. Plan II provides a refundable earned income credit of 14 percent (up to \$1,000) to all earners and increases the credit by 4 percentage points for each person in the tax unit above one (with a phaseout between \$10,000 and \$20,000 of adjusted gross income).

[d] Range within which the 13-percentage point reduction in the first bracket is phased out. Top limit of the range increases or decreases by \$11,480 for each personal exemption.

TABLE 8—EFFECTIVE FEDERAL INDIVIDUAL INCOME TAX RATES UNDER THE TAX REFORM ACT OF 1986 AND UNDER TWO COMPREHENSIVE INCOME TAXES, 1990[a]

	Percent		
Population Percentile[b]	Tax Reform Act of 1986	Comprehensive Income Tax[c]	
		Plan I	Plan II
1st Decile[d]	−0.9	−0.4	−1.3
2nd Decile	0.0	0.1	−0.8
3rd Decile	1.9	2.0	0.7
4th Decile	4.5	4.6	3.3
5th Decile	6.2	6.2	5.0
6th Decile	7.4	7.4	6.1
7th Decile	8.4	8.5	7.0
8th Decile	9.4	9.5	7.8
9th Decile	11.1	10.9	9.1
10th Decile	16.5	16.5	18.9
Top 5 Percent	18.0	18.1	22.3
Top 1 Percent	20.7	20.8	30.5
All Deciles[e]	11.1	11.1	11.1

Source: Congressional Budget Office.

[a] Assumes corporate income is borne by capital income.

[b] Arrayed by comprehensive income, which includes all transfer payments, the employer share of payroll taxes, and the corporate income tax (allocated to capital income).

[c] For tax rates under the two plans, see Table 7.

[d] Excludes families with zero or negative incomes.

[e] Includes families with zero or negative incomes.

able income above $250,000 would accomplish this objective (see Plan II, Table 7).[49]

Another change to increase progressivity would be to increase the earned-income credit for low-income families. Today, the credit is the same for all families, regardless of the number of children. The credit would be more effective in combatting poverty if it increased with family size. For example, the current 14-percent credit could be maintained for families with one child and four percentage points, or roughly $250, could be added for each additional child. With this modification, the earned income credit would increase the likelihood that a family with several children could earn enough to remain outside the welfare system.

Table 8 reports the average effective federal income tax rates by population deciles under the schedule that restores overall progressivity to 1977 levels (Plan II) and under the schedule that matches 1990 distribution with a broadened base (Plan I). The average effective rates in Plan II are lower than those under present law in the bottom nine deciles and higher in the top decile. For the top 1 percent of the family units, the average effective rate rises from 21 percent to 30 percent, which cannot be regarded as punitive.

I recognize that few people would go as far as I would in broadening the tax base. But that does not mean that the objective of greater progressivity must be abandoned. Even if there were no additional base-broadening between now and 1990, the same degree of progressivity that prevailed in 1977 could be achieved with rates ranging from 7 percent at the bottom to 56 percent at the top.

Conclusion

I conclude that there is no good reason for the disenchantment of economists with the income tax. The main rival of the income tax —the consumption expenditure tax— is distinctly inferior on theoretical as well as practical grounds. The endowment or lifetime perspective of the expenditure tax is indefensible in a world of financial, political, and family instability. The transition problems in moving from an income tax to an expenditure tax are extremely difficult. There is also a danger that the substitution of an expenditure tax for the income tax would greatly increase the concentration of wealth. Moreover, the public regards income, not expenditure, as the best index of ability to pay, and it would be unwise to abandon this familiar and widely approved basis of taxation.

The 1986 reforms have greatly improved the federal income tax by broadening the base and lowering rates. But the progressivity of the federal tax system has been declining for the last two decades. As a result, the distribution of before-tax income, which has been growing more unequal in the 1980s, has become even more unequal on an after-tax basis. I have suggested that the goal of tax policy should be to restore the progressivity of the income tax at least to its level in the mid-1970s.

The 1986 tax reform went a long way toward comprehensive income taxation, but much more can be done to enlarge the tax base and to remove the preferences for capital income. Among the more urgent base-broadening reforms are the inclusion in taxable income of capital gains transferred by gift or at death, elimination of the tax exemption for newly issued state-local securities, taxation of employee fringe benefits, treatment of Social Security benefits like private pensions, reduction of the tax subsidy for home owners, pruning of the personal deductions, and withholding on interest and dividends. To correct the measurement of capital income for inflation, asset prices should be adjusted for changing prices in order to convert nominal to real incomes for tax purposes. The two-earner deduction and income averaging should be restored to reduce the marriage penalty and equalize the treatment of fluctuating and stable incomes.

---

[49] Between 1977 and 1990 the weight of the relatively regressive payroll tax in federal revenues increased. Thus, the income tax must be more progressive in 1990 than it was in 1977 to restore overall progressivity to its 1977 level.

A comprehensive income tax along these lines would permit further rate reductions throughout the income scale if the degree of progression enacted in 1986 were to be retained. However, the progressivity of the federal tax system has declined since the mid-1970s, even after taking into account the effect of the 1986 act. To restore the degree of progressivity of the mid-1970s, the rate of graduation of the tax rates would need to be increased. I estimate that this can be accomplished with a rate schedule ranging from 4 percent at the bottom to 48 percent at the top of the taxable income scale—a moderate schedule of rates by any standard.

It is clear from this analysis that the revenue potential of the income tax has not been exhausted in this country. Even if the base is not broadened, the income tax can be used to raise considerable additional revenues in order to eliminate the recurring federal deficits. Each percentage-point increase in the individual and corporate income tax rates would bring in about $30 billion in 1994, so that three points would come close to balancing the overall budget in that year. A top individual income tax rate of 31 percent and a corporate rate of 37 percent cannot be regarded as punitive or harmful to economic incentives.

What is inappropriate in my view would be to introduce a value-added tax, as some are suggesting. The value-added tax is regressive and imposes unnecessarily heavy burdens on the lower income classes. With tax rates as low as they are today, more revenues should come from the income tax, the tax paid by those who have the ability to pay. In view of the recent reductions in the progressivity of the federal tax system, it would be unconscionable to enact the distinctly inferior alternative of a value-added tax.

REFERENCES

Aaron, Henry J., "Politics and the Professors Revisited," *American Economic Review*, May 1989, *79*, *Papers and Proceedings* 1988, 1–15.

______ **and Galper, Harvey**, *Assessing the Income Tax*, Washington, DC: Brookings Institution, 1985.

Andrews, William D., "A Consumption-Type or Cash Flow Personal Income Tax," *Harvard Law Review*, April 1974, *87*, 1113–88.

Atkinson, Anthony B. and Sandmo, Agnar, "Welfare Implications of the Taxation of Savings," *Economic Journal*, September 1980, *90*, 529–49.

______ **and Stiglitz, Joseph E.**, *Lectures on Public Economics*, New York: McGraw-Hill, 1980.

Auerbach, Alan J. and Kotlikoff, Laurence J., *Dynamic Fiscal Policy*, London: Cambridge University Press, 1987.

______ **and Poterba, James M.**, "Why Have Corporate Tax Revenues Declined?," *Tax Policy and the Economy*, Lawrence H. Summers, ed., Vol. 1, 1987.

Ballard, Charles L., Fullerton, Don, Shoven, John B. and Whalley, John, *A General Equilibrium Model for Tax Policy Evaluation*, Chicago: University of Chicago Press, 1985.

Blum, Walter J. and Kalven, Harry, "The Uneasy Case for Progressive Taxation," *University of Chicago Law Review*, Spring 1952, *19*, 417–520.

Boskin, Michael J. "Taxation, Saving, and the Rate of Interest," *Journal of Political Economy*, April 1978, *86*, pt. 2, S3–S27.

Bosworth, Barry P., *Tax Incentives and Economic Growth*, Washington, DC: Brookings Institution, 1984.

Bradford, David F., "The Case for a Personal Expenditure Tax," *What Should Be Taxed: Income or Expenditure?*," Washington, DC: Brookings Institution, 1980, 75–113.

______, *Untangling the Income Tax*, Cambridge: Harvard University Press, 1986.

Break, George F., "Income Taxes and Incentives to Work: An Empirical Study," *American Economic Review*, September 1957, *5*, 529–49.

Burtless, Gary, "The Supply-Side Legacy of the Reagan Years: Response of Labor Supply," unpublished paper, July 3, 1989.

______ **and Haveman, Robert**, "Taxes Transfers, and Labor Supply: The Evolving Views of U.S. Economists," *The Relevance of Public Finance for Policy-Making*, Detroit: Wayne State University Press.

Clotfelter, Charles, "The Impact of Tax Reform on Charitable Giving: A 1989 Perspective," paper prepared for a Conference on the Economic Effects of the Tax Reform Act of 1986, Ann Arbor, MI, November 10–11, 1989.

Congressional Budget Office, *The Changing Distribution of Federal Taxes: 1975–1990*, October 1987.

______, *The Changing Distribution of Federal Taxes: A Closer Look at 1980*, July 1988.

______, *Reducing the Deficit: Spending and Revenue Options*, February 1989.

Economic Report of the President, Washington, DC: U.S. Government Printing Office, February 1982.

Courant, Paul N. and Gramlich, Edward M., "Tax Reform: There Must be a Better Way," *National Policy Papers*, Washington, DC: National Policy Exchange, November 1982.

Danziger, Sheldon, Gottschalk, Peter and Smolensky, Eugene, "How the Rich Have Fared, 1973–87," *American Economic Review, Papers and Proceedings* 1988, May 1989, *79*, 310–14.

Evans, Owen J., "Tax Policy, the Interest Elasticity of Saving, and Capital Accumulation: Numerical Analysis of Theoretical Models," *American Economic Review*, June 1983, *73*, 398–410.

Feldstein, Martin, "The Welfare Cost of Capital Income Taxation," *Journal of Political Economy*, April 1978, *86*, S29–S51.

Fichtenbaum, Rudy, and Shahidi, Hushang, "Truncation Bias and the Measurement of Income Inequality," *Journal of Business and Economic Statistics*, July 1988, *5*, 335–37.

Fisher, Irving, "Income in Theory and Income Taxation in Practice," *Econometrica*, January 1937, *5*, 1–55.

Galper, Harvey, Lucke, Robert and Toder, Eric, "A General Equilibrium Analysis of Tax Reform," *Uneasy Compromise: Problems of a Hybrid Income-Consumption Tax*, in Henry J. Aaron, Harvey Galper, and Joseph A. Pechman, eds., Washington, DC: Brookings Institution, 1988, 59–114.

Goode, Richard, *The Individual Income Tax*, rev. ed., Washington, DC: Brookings Institution, 1976.

______, "Long-Term Averaging for Income Tax Purposes," *The Economics of Taxation*, Washington, DC: Brookings Institution, 1980, 159–78.

Gravelle, Jane G., "Assessing a Value-Added: Efficiency and Equity," *Tax Notes*, March 7, 1988, 1117–23.

______, "Income, Consumption, and Wage Taxation in a Life Cycle Model: Separating Efficiency from Redistribution," unpublished paper, June 1989.

Hall, Robert E., and Rabushka, Alvin, *The Flat Tax*, Stanford, CA: Hoover Institution Press, 1985.

Hausman, Jerry A., "Taxes and Labor Supply," *Handbook of Public Economics*, Alan J. Auerbach and Martin Feldstein, eds., Amsterdam: North-Holland, 1987, Vol. 1, 213–63.

______ **and Poterba, James M.,** "Household Behavior and the Tax Reform Act of 1986," *Journal of Economic Perspectives*, Summer 1987, *1*, 101–20.

Holland, Daniel M., "The Effect of Taxation on Effort: Some Results for Business Executives," *Proceedings of the Sixty-Second Annual Conference on Taxation, 1969*, Columbus OH: National Tax Association, 1970, 428–517.

Hubbard, R. Glen and Judd, Kenneth L., "Liquidity Constraints, Fiscal Policy, and Consumption," *Brookings Papers on Economic Activity*, *1:1986*, 175–208.

Kenadjian, Bardj, Internal Revenue Service, "Gross Tax Gap Trends According to New IRS Estimates, Income Years 1973–1992," *Statistics of Income Bulletin*, Summer 1988, *8*, 23–33.

Kaldor, Nicholas, *An Expenditure Tax*, London: Allen & Unwin, 1955.

Kay, John A. and King, Mervyn A., *The British Tax System*, 3rd ed., Oxford: Oxford University Press, 1983.

Laffer, Arthur B., "Government Exactions and Revenue Deficiencies," *Cato Journal*, Spring 1981, *1*, 1–21.

Levy, Frank, *Dollars and Dreams: The Changing American Income Distribution*, New York: Norton, 1988.

McLure, Charles E., Jr., "Where Tax Reform Went Astray," *Villanova Law Review*, 1987, *31*, 1619–63.

______, **Mutti, John, Thuroni, Victor, and Zodrow, George,** *The Taxation of Income from Business and Capital in Colombia*, Colombia: Direccion General de Impuestos Nacionales, Ministerio de Hacienda y Credito Publico, 1988.

Meade, James, et al., *The Structure and Reform of Direct Taxation*, London: Allen & Unwin, 1978.

Meyer, Lawrence H., ed., *The Supply-Side Effects of Economic Policy*, The Hague: Kluwer Nijhoff Publishing, 1981.

Mieszkowski, Peter, "The Advisability and Feasibility of an Expenditure Tax System, *The Economics of Taxation*, Henry J. Aaron and Michael J. Boskin, eds., Brookings Institution, 1980, 179–202.

Mill, John Stuart, *Principles of Political Economy*, W. J. Ashley, ed., London: Longmans, Green, 1929, Bk 2, ch. 2, sec. 4.

Musgrave, Richard A., *The Theory of Public Finance: A Study in Public Economy*, New York: McGraw-Hill, 1959.

______, "ET, OT, and SBT," *Journal of Public Economics*, July-August 1976, *6*, 3–16.

______, *Strengthening the Progressive Income Tax: The Responsible Answer to America's Budget Problem*, Washington, DC: Economic Policy Institute, 1989.

Pechman, Joseph A., ed., *Comprehensive Income Taxation*, Washington, DC: Brookings Institution, 1977.

______, **ed.,** *What Should be Taxed: Income or Expenditure?*, Washington, DC: Brookings Institution, 1980.

______, *Who Paid the Taxes, 1966–85?*, Washington, DC: Brookings Institution, 1985.

______, *Who Paid the Taxes, 1966–85?, Revised Tables*, Washington, DC: Brookings Institution, 1987.

______, (1987a) "Tax Reform: Theory and Practice," *Journal of Economic Perspectives*, Summer 1987, *1*, 11–28.

______, (1987b) *Federal Tax Policy*, 5th ed., Washington, DC: Brookings Institution, 1987.

______, **ed.,** *World Tax Reform: A Progress Report*, Washington, DC: Brookings Institution, 1988.

______, *Tax Reform, The Rich and the Poor*, Washington, DC: Brookings Institution, 1989.

Pechman, Joseph A. and Okner, Benjamin A., *Who Bears the Tax Burden?*, Washington: Brookings Institution, 1974.

Raboy, David G., *Essays in Supply Side Economics*, Washington, DC: Institute for Research in the Economics of Taxation, 1982.

Reagan, Ronald, *Public Papers of the Presidents of the United States, 1981*, Washington, DC: U.S. Government Printing Office, 1982.

Rosen, Harvey S., *Public Finance*, 2nd ed., Homewood, IL: Irwin, 1988.

______, "Does Progressive Taxation Survive Optimal Theory" unpublished paper, 1988.

Simons, Henry C., *Federal Tax Reform*, Chicago: University of Chicago Press, 1950.

______, *Personal Income Taxation: The Definition of Income as a Problem of Fiscal Policy*, Chicago: University of Chicago Press, 1938.

Slemrod, Joel, "Optimal Taxation and Optimal Tax Systems," *Journal of Economic Perspectives*, (forthcoming).

Stern, Nicholas, "The Theory of Taxation and Optimum Income Taxation: An Introduction," *The Theory of Taxation for Developing Countries*, David Newberry and Nicholas Stern, eds., Washington, DC: Oxford University Press for the World Bank, 1987, 22–59.

Steuerle, C. Eugene, *Taxes, Loans and Inflation: How the Nation's Wealth Becomes Misallocated*, Washington, DC: Brookings Institution, 1985.

Summers, Lawrence H., "Capital Taxation and Accumulation in a Life-Cycle Growth Model," *American Economic Review*, September 1981, *71*, 533–44.

U.S. Department of the Treasury, *Blueprints for Basic Tax Reform*, Washington, DC: U.S. Government Printing Office, 1977.

Young, H. Peyton, "Progressive Taxation and Equal Sacrifice," *American Economic Review*, March 1990, *80*, 253–66.

Equilibrium Political Budget Cycles

By KENNETH ROGOFF*

Political business cycle theories generally rely on nominal rigidities and voter myopia. This paper offers an equilibrium theory which preserves some basic insights from earlier models, though with significant refinements. The "political budget cycle" emphasized here is in fiscal policy rather than output and inflation; it arises via a multidimensional signal process. One can consider the welfare implications of proposals to mitigate the cycle, and the effects of altering the electoral structure. (JEL 131)

Economists and political scientists have long been intrigued by the coincidence of elections and economic policy cycles.[1] During election years, governments at all levels often engage in a consumption binge, in which taxes are cut, transfers are raised, and government spending is distorted toward projects with high immediate visibility. The proximate cause of the "political budget cycle" does not seem difficult to identify. Any incumbent politician, regardless of his ideological stripes, wants to convince voters that he is doing an efficient job running the government. The deeper question is why rational voters might allow their expectations about postelection performance to be influenced by preelection budget antics.[2]

In this paper, I offer a dynamic, multidimensional signaling model in which both voters and politicians are rational, utility-maximizing agents. A political budget cycle arises here due to temporary information asymmetries about the incumbent leader's "competence" in administering the public goods production process. The incumbent leader has an incentive to bias preelection fiscal policy toward easily observed consumption expenditures, and away from government investment. In equilibrium, however, voters can deduce the leader's current competency by the degree to which he distorts tax and expenditure policies.[3]

Perhaps the most important reason for trying to develop a fully articulated equilibrium model of political budget cycles is to enable one to analyze the welfare implications of alternative electoral regimes, and of various proposals for tempering election year budget distortions.[4] For example, the popu-

---

*Economics Department, University of California at Berkeley, CA 94720. This research has been supported by the National Science Foundation under grant no. SES-87-20800, and by the Alfred P. Sloan Foundation, Much of the work was conducted while the author was on leave as a National Fellow at the Hoover Institution. I am grateful to Marco Terrones and Maurice Obstfeld for helpful comments on an earlier draft.

[1] Important contributions to the modern literature on political business cycles include William D. Norhaus (1975), Assar Lindbeck (1976), and Edward R. Tufte (1978). For a broad survey of the more recent literature on politics and macroeconomic policy, see Alberto Alesina (1988); see also Thomas D. Willet (1989).

[2] Kenneth Rogoff and Anne Sibert (1988) show that political budget cycles can be given an equilibrium signaling interpretation. Their model is not sufficiently articulated, however, to address the normative issues raised here.

[3] The rationale for political budget cycles here is very different from the one underlying the "partisan" models of Douglas A. Hibbs (1977) and Alesina (1987). In partisan models, two parties with very different preferences over inflation and unemployment compete for office. Consequently, private nominal wage setters have great difficulty predicting postelection monetary policy. For empirical evidence on this approach, see Alesina (1988), (1989), Daniel Cohen (1988) and Steven M. Sheffrin (1988). In principle, it should be possible to generalize the present analysis to incorporate partisan factors.

[4] A number of authors have previously addressed normative aspects of political business cycles; see Lindbeck (1976), Henry Chappell and William Keech (1983),

P, 5779

lar perception is that political budget cycles are a bad thing. But a central conclusion here is that they may be a socially efficient mechanism for diffusing up-to-date information about the incumbent's administrative competence. Efforts to curtail the cycle can easily reduce welfare, either by impeding the transmission of information or by inducing politicians to select more socially costly ways of signaling.[5]

In Section I, I present the model, including the constitutionally constrained election structure. Section II gives the equilibrium under full information, and Sections III–V characterize the fiscal policy distortions which occur under asymmetric information. There are multiple sequential equilibria to the model, but after applying some standard refinements, one obtains a unique equilibrium. In Section VI, I explore some possible approaches to mitigating the cycle, including a constitutional limit on the legislature's ability to undertake new fiscal initiatives directly prior to elections. One interesting alternative electoral structure, common to many countries, gives the incumbent the option of calling for an early election. Finally, I consider whether there are ways for society to channel preelection signaling into dimensions that primarily impact on the incumbent and not on society at large. In the conclusions, the predictions of the equilibrium political budget cycle theory are compared with those of its Keynesian predecessor, the political business cycle theory.

I. The Model

A. *The Preferences of a Representative Citizen*

The economy is composed of a large number of (*ex ante*) identical citizens, each of whom derives utility both from public goods and from a private consumption good. The representative voter is concerned with the expected value of his utility function, $E_t^P(\Gamma_t)$, where subscripts denote time, E^P denotes expectations based on the general public's information set, and

$$(1) \quad \Gamma_t = \sum_{s=t}^{T} \left[U(c_s, g_s) + V(k_s) + \eta_s \right] \beta^{s-t}.$$

In equation (1), c is a representative citizen's consumption of the private good, g is the public "consumption" good (per capita), and k is the public "investment" good (per capita). (Population will be held constant throughout). U and V are both regular strictly concave functions, with $U_1, U_2, V' > 0$. In addition to the usual Inada conditions, I make the further assumption that $\lim_{k\to 0} V(k) = -\infty$ [for example, $V(k) = \log k$]. This condition is sufficient to ensure an interior solution in the asymmetric information case. $\beta < 1$ is the representative citizen's discount rate, and T is his time horizon, which may be infinite. The term η is a random shock, which will later be identified with non-pecuniary leader-specific factors such as the leader's looks.

B. *Technology*

At the beginning of each period, all citizens exogenously receive y units of a nonstorable good, which can either be privately consumed or used as an input into the production of public goods. Lump-sum taxes in period t are given by τ_t, so that

$$(2) \quad c_t = y - \tau_t.$$

In addition to taxes, the production of public goods also requires a (single) "leader" whose administrative competency is indexed by ε. A competent administrator (high ε) is able to provide a given level of public goods at a lower level of taxes than an incompetent one can. Specifically, I assume that the pub-

---

Keech and Carl Simon (1985), Alex Cukierman and Allan H. Meltzer (1986), and Willett (1989). None of these analyses, however, are based on fully specified equilibrium models.

[5] Tufte (1978, p. 149) also suggests that political business cycles may have socially beneficial aspects. He argues that the government tends to distribute income more equitably prior to elections than at other times.

lic goods production function takes the form[6]

$$g_t + k_{t+1} = \tau_t + \varepsilon_t. \tag{3}$$

Note that whereas the relative cost of producing g and k is unity, the timing of their production differs. To have the public "investment" good k in period $t+1$, the government must invest in period t.

C. *Stochastic Structure*

All agents are capable of serving as the country's leader. However, at any point in time they differ according to their innate administrative ability. For each agent i, (potential) leadership competency evolves according to the serially correlated stochastic process

$$\varepsilon_t^i = \alpha_t^i + \alpha_{t-1}^i, \tag{4}$$

where each α is an independent drawing from a Bernoulli distribution with $\rho \equiv \text{prob}(\alpha = \alpha^H)$ and $1 - \rho \equiv \text{prob}(\alpha = \alpha^L)$, $\alpha^H > \alpha^L > 0$. The α shocks are independent across agents as well as across time. One reason why competency might realistically be thought to vary across time is that leadership abilities well-suited to dealing with the one set of historical circumstances may become outmoded as the problems facing the country change. Also, even if the same leader stays in power, there may be turnover among his key advisors.

In this model, competency is not a choice variable for a leader but an individual characteristic. One may think of competency as administrative IQ.[7] The assumption that competency follows a first-order moving average process simplifies the analysis below considerably, by effectively breaking structural links between elections. However, the main qualitative results here should carry over to more general stochastic processes.

In addition to the competency shock, each agent i experiences a "looks" shock η, which also follows a moving average process:

$$\eta_t^i = q_t^i + q_{t-1}^i, \tag{5}$$

where each q is a continuously distributed i.i.d. random variable on $[-\bar{q}, \bar{q}]$, with q_t^i and q_s^j independent for all $s \neq t$, $i \neq j$. The random variable η is intended to capture factors relevant to an agent's leadership ability but uncorrelated with his competence in administering the public goods production function; for example, his "looks." Neither ε^i nor η^i matter for anything when agent i is a private citizen. Throughout, whenever ε, η, α, or q are written without a superscript, they refer to the incumbent.

D. *The Leader's Utility Function*

The country's leaders are drawn from the ranks of ordinary citizens and as such, they derive utility from public and private consumption goods in the same way as other citizens. However, because the position of chief administrator is considered a great honor, the leader receives additional "ego rents" of X per period in office. Thus, for an incumbent leader, expected utility is given by[8]

$$E_t^I(\Gamma_t) + \sum_{s=t}^{T} \beta^{s-t} X \pi_{s,t}, \tag{6}$$

where I denotes the incumbent and Γ is given by (1). E_t^I denotes expectations based on the incumbent's information set at time t, and $\pi_{s,t}$ is the incumbent's time t estimate of his probability of being in office in period s; π will be derived later on. Prospective ego rents (X) do not enter explicitly into the objective function of an ordinary citizen, $E^{P}(\Gamma)$, only because the population is sufficiently large that the probability of his ever being elected is infinitesimal.

---

[6] The analysis would be similar in most respects if ε entered the production function multiplicatively, either multiplying $g+k$ or k alone. There would be some differences since a change in ε then has price effects as well as income effects, but the welfare results in Section VI would not be affected.

[7] It is tempting to stretch the paradigm here to interpret competency as reflecting the efficacy of a particular political party's general philosophical approach toward managing the government.

[8] Implicit in (6) is the assumption that the leader cares just as much about his own "looks" shock as private agents do. The results below would be the same, however, if η did not enter the leader's objective function. Another assumption implicit in (6) is that the leader is not legally allowed to tax himself differently from other individuals; there is no graft.

I have motivated the leader's utility function without any appeal to altruism. But it should later be apparent that the analysis would be similar in most respects under a more generous interpretation of the leader's aims. Specifically, one can interpret equation (6) as saying that the leader puts some weight on social welfare, Γ, and some weight on the rents he receives from being in office.

E. *The Structure of Elections*

In order to determine which citizen is awarded the honor of administering the production of public goods, the country's constitution specifies that elections be held every other period. The incumbent leader is allowed to run an indefinite number of times,[9] whereas the opposition candidate is chosen at random from the rest of the population. Note that under the information structure specified below, the fundamental difference between the incumbent leader and his opponent is that the public can infer something about the incumbent's most recent competency shock, but it has no way of inferring anything about the opponent's competency. For voters, the choice is essentially between reelecting the incumbent or selecting an agent from the population at large, all of whom appear identical *ex ante*.[10]

F. *The Information Structure and the Timing of Events*

Voters observe taxes τ_t and government consumption spending g_t contemporaneously, and employ this information to form *inferences* concerning government investment spending k_{t+1} and the incumbent's competency shock α_t. However, they cannot directly confirm these inferences until the following period. In period $t+1$, the government's period t investment comes on line and voters also directly observe α_t. Thus the incumbent has a temporary information "advantage" over voters in the sense that he sees his competency shock contemporaneously. I use the word "advantage" in quotation marks because it will turn out that *in equilibrium* voters are always able to deduce the incumbent's private information.

The information structure assumed here is plausible since it is costly for an individual to closely monitor and evaluate a government's performance. Moreover, there is little private incentive for an individual to undertake such monitoring since in equilibrium, he can infer α_t costlessly using his information on g_t and τ_t. Taxes and government consumption spending are variables which individuals need to know and can observe relatively easily. On the other hand, if k represents investment in national defense, there may be national security reasons for not making it public. More broadly, k may be thought of as vesting of public pension funds, off-budget loan guarantees, or any type of government expenditure whose effects are only observed by the representative voter with a lag.

Of course, if some group were able to monitor the government and *credibly* transmit information in a way that would not be too costly for the average citizen to process, then there would be no political budget cycle in the analysis below. Clearly, neither the opposition candidate nor the incumbent can provide this service, since their statements cannot be trusted. The results below should go through in a more general setting in which some voters monitor α_t, as long as there is a sufficient pool of uninformed voters.

---

[9] If the incumbent can only run for reelection a finite number of times, then the model predicts that there will be no political budget cycle in the last period. One can also interpret the model along the lines of Rogoff and Sibert (1988), in which electoral competitions match two political parties.

[10] The stochastic structure of the model is consistent with Ray C. Fair's (1978) finding that for U.S. presidential elections, voters do not take into account the opposition party's economic performance when last in power; see also Sam Peltzman (1987).

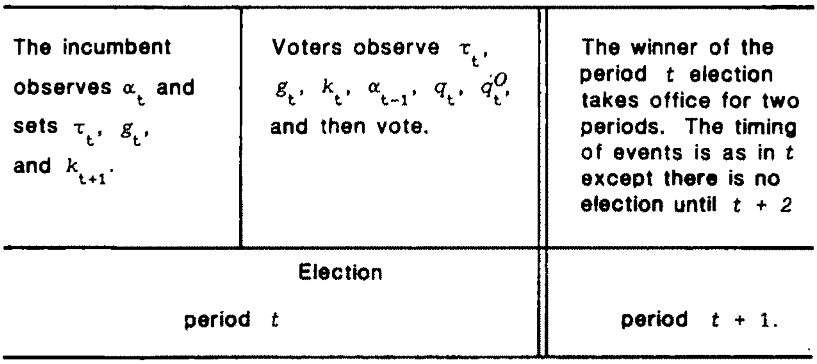

The incumbent observes α_t and sets τ_t, g_t, and k_{t+1}.
Voters observe τ_t, g_t, k_t, α_{t-1}, q_t, q_t^O, and then vote.
The winner of the period t election takes office for two periods. The timing of events is as in t except there is no election until $t + 2$
Election
period t
period $t + 1$.

FIGURE 1. THE TIMING OF EVENTS

The public, of course, has no way of inferring α_t^O, where "O" superscripts denote the opponent. All voters know about the opponent is the probability distribution of α. (The incumbent has no way of knowing α_t^O either until he actually tries his hand at running the government.) Prior to voting, voters do observe both q_t and q_t^O, the "looks" shocks.

The incumbent must set g and τ prior to observing the q's. The rationale for this assumption is that it takes time for the government to collect taxes and to make purchases. The q shocks, on the other hand, might capture information revealed in election-eve debates or uncertainty about a last-minute scandal concerning one of the candidates. In a slightly different version of the model in which voters have heterogeneous preferences over looks, q can represent uncertainty about election-day weather, and thus about the composition of voters who come to the polls.[11] Figure 1 illustrates the timing of events.

In deciding his vote, the representative voter compares his expected utility under each of the two candidates. If $\nu = 1$ denotes a vote for the incumbent and $\nu = 0$ a vote for his opponent, then

$$(7) \quad \nu_t = \begin{cases} 1 & \text{if } E_t^P(\Gamma_{t+1}) \geq E_t^P(\Gamma_{t+1}^O) \\ 0, & \text{otherwise.} \end{cases}$$

II. Equilibrium Under Full Information

Before proceeding, it is useful to analyze the equilibrium which would arise if voters could directly observe α_t prior to voting. In this case, the incumbent's preelection fiscal policy cannot possibly affect voter's expectations abut his postelection competency, and thus can have no effect on his chances of remaining in office. With the π terms in (6) exogenous, the incumbent's decision problem becomes equivalent to maximizing the welfare of the representative agent. Given the simple production and storage technology, this problem can be broken down into a sequence of static maximization problems:

$$(8) \quad \max_{\tau_t, c_t, g_t, k_{t+1}} U(c_t, g_t) + \beta V(k_{t+1}), \qquad \forall\, t \geq T$$

subject to (2), (3), and $k, c, g \geq 0$; $k_{T+1} = \bar{k}$.[12]

---

[11] The analysis below is quite similar when voters have heterogeneous tastes concerning "looks," except that elections are no longer unanimous.

[12] Note that τ is allowed to take on negative values (the government can make net transfers). This assumption is not qualitatively important and the analysis is easily generalized to the case where taxes are constrained to be nonnegative.

It is convenient to rewrite the above maximization problem by substituting (2) and (3) into (8):

$$(9)\quad \max_{\tau, g} W(g,\tau,\varepsilon) \equiv U(y-\tau, g) + \beta V(\tau+\varepsilon-g)$$

$$\text{s.t. } g, y-\tau, \tau+\varepsilon-g \geq 0.$$

(Time subscripts will henceforth be omitted where the meaning is obvious.) The first-order conditions for an interior solution to (9) imply

$$(10)\quad U_1(y-\tau,g) = U_2(y-\tau,g),$$

$$(11)\quad U_1(y-\tau,g) = \beta V'(\tau+\varepsilon-g).$$

One can readily confirm that there is a unique $[g^*(\varepsilon), \tau^*(\varepsilon)]$ which satisfies (10) and (11), and that this point is a global maximum. (Note that U and V are strictly concave and that the constraint set is convex.) Clearly

$$W^*(\varepsilon) = W^*[g^*(\varepsilon), \tau^*(\varepsilon), \varepsilon]$$

is strictly increasing in ε and, if all goods are normal, then $c^*(\varepsilon)$, $g^*(\varepsilon)$, and $k^*(\varepsilon)$ are also increasing. By (2), $\tau^*(\varepsilon)$ must be decreasing in ε.

If t is an election period, then by equations (1), (4), (5), (7), and (9), the incumbent will be reelected ($\nu = 1$) if

$$(12)\quad E_t^P[W^*(\varepsilon_{t+1})] - E_t^P[W^*(\varepsilon_{t+1}^O)] + q_t - q_t^O \geq 0.$$

Because ε and η follow first-order moving average processes, voters' expected utility is the same under either candidate for periods $t+2$ and beyond, and thus only expectations over $t+1$ enter into (12). (Recall from Figure 1 that voters observe the q shocks prior to the election.)

If voters directly observe the incumbent's most recent competency shock prior to voting, then the first term in (12) is given by

$$(13)\quad E_t^P[W^*(\varepsilon_{t+1}) | \alpha_t = \alpha^i] \equiv \Omega^i = \rho W^*(\alpha^i + \alpha^H) + (1-\rho) W^*(\alpha^i + \alpha^L); \qquad i = H, L.$$

Voters have no observations on the opponent's competency; hence

$$(14)\quad E_t^P[W^*(\varepsilon_{t+1}^O)] \equiv \Omega^O = \rho^2 W^*(2\alpha^H) + 2\rho(1-\rho) W^*(\alpha^H + \alpha^L) + (1-\rho)^2 W^*(2\alpha^L).$$

Clearly, $\Omega^H > \Omega^O > \Omega^L$.

III. Voters' and Leaders' Optimization Problems Under Asymmetric Information

I now return to the asymmetric information structure summarized in Figure 1. Although the public cannot observe α_t until period $t+1$, they can form "beliefs" about α_t given their observations on g_t and τ_t. These beliefs can be parameterized as $\hat{\rho}(g,\tau)$, where $\hat{\rho}$ is the probability weight the public attaches to the possibility that $\alpha_t = \alpha_H$. (Since α_{t-1} is a fixed, known parameter throughout this section, I abbreviate $\hat{\rho}(\alpha_t, \tau_t; \alpha_{t-1})$ as $\hat{\rho}(\alpha, \tau)$).

We will initially focus on the final election period, $t = T-2$. Since the winner will not be running for reelection, he has no incentive to distort fiscal policy in periods $T-1$ or T. Thus $E_t^P[W(\varepsilon_{T+1})] = E_t^P[W^*(\varepsilon_{T+1})]$ if the incumbent wins and similarly for his opponent. By equations (12)–(14), if voters have priors $\hat{\rho}(g,\tau)$, the incumbent will be reelected ($\nu = 1$) if

$$(15)\quad \hat{\rho}\Omega^H + (1-\hat{\rho})\Omega^L - \Omega^O + q - q^O \geq 0.$$

The incumbent does not know $q - q^O$ when setting his election-year fiscal policy. However, for any choice of (g, τ), he can infer $\hat{\rho}(g, \tau)$ and thus calculate the probability that $q - q^O$ will be high enough for him to win:

$$(16)\quad \pi[\hat{\rho}(g,\tau)] \equiv E^I(\nu | g, \tau) = 1 - G[\Omega^O - \hat{\rho}\Omega^H - (1-\hat{\rho})\Omega^L],$$

where G is the probability distribution function of $q - q^O$. The possibility for signaling arises here because there is a limit to how much an incumbent would be willing to distort fiscal policy in order to fool the public about his competency. As a representative agent, he too cares about the mix of consumption and investment; see (6) above.

It is convenient to define $\varepsilon^H = \alpha_{t-1} + \alpha^H$, and $\varepsilon^L = \alpha_{t-1} + \alpha^L$. The incumbent will be described as a "type H" (or "competent type") if $\varepsilon_t = \varepsilon^H$, and a "type L" (or "incompetent type") if $\varepsilon_t = \varepsilon^L$. Using equations (1), (4)–(6), (9), and (16), one can then write an incumbent of type i's maximization problem as

$$(17)\quad \max_{g,\tau} Z[g, \tau, \hat{\rho}(g,\tau), \varepsilon^i]$$

$$\text{s.t. } g, y - \tau, \tau + \varepsilon^i - g \geq 0; \quad i = H, L,$$

where

$$(18)\quad Z[g, \tau, \hat{\rho}(g,\tau), \varepsilon^i] \equiv \chi^i \pi[\hat{\rho}(g,\tau)] + W(g, \tau, \varepsilon^i)],$$

$$(19)\quad \chi^i \equiv \beta[X(1+\beta) + \Omega^i - \Omega^O].$$

The first term on the RHS of (18) is the incumbent's expected chance of winning, π, multiplied by his surplus from winning, χ^i. This surplus is broken down in (19), where the term $X(\beta + \beta^2)$ captures the discounted ego rents for the two postelection periods, and the term $\beta(\Omega^i - \Omega^o)$ is the amount by which the representative citizen's expected utility is higher if the incumbent wins instead of his opponent. I assume $\chi^L > 0$.

Two features distinguish the objective function of a competent type from that of an incompetent type. First, the competent type knows that expected social welfare will be higher if he is reelected than if his unknown opponent wins. The second difference is that for any (g, τ), a type H is investing $\alpha^H - \alpha^L$ more units into k_{t+1} than a type L is, by equation (3). An important implication is that since $V'' < 0$, a type H can cut back on government investment at lower *marginal* cost that can a type L.

IV. Sequential Equilibria

The interaction between incumbent politicians and rational voters here can be viewed as a multidimensional signaling problem, with g and τ as signals of the incumbent's (contemporaneously) unobserved competency. As is typically the case in such models, there is a multiplicity of sequential equilibria, including both *separating* and *pooling* equilibria. In a separating equilibrium, the incumbent's choice of fiscal policy perfectly reveals his competency type. In a pooling equilibrium, the incompetent type might mimic the competent type. However, by requiring that voters' beliefs reflect a certain minimal level of sophistication (by excluding "dominated" strategies),[13] it is possible to rule out all but one of the separating equilibria. By further refining the equilibrium concept, using the "intuitive" criterion of In-Koo Cho and David Kreps (1987), one can also rule out pooling equilibria. In the unique equilibrium which survives both refinements, competent types set taxes too low and government spending too high before elections, whereas incompetent types pursue their full information policy. On average, there is a political budget cycle.[14]

In the main text, I will restrict attention to equilibria in pure strategies. For $i = L, H$, let

---

[13] See Hervé Moulin (1981). The general approach here draws on Paul Milgrom and John Roberts (1986), and Kyle Bagwell and Garey Ramey (1988).

[14] The analysis can be generalized to allow for a continuum of types along the lines of Rogoff and Sibert (1988). In their analysis, very competent and very incompetent types distort the least.

(g^i, τ^i) describe a strategy for the incumbent leader, and let $\nu[\hat{\rho}(g,\tau), q - q^O]$ describe a strategy for voters. Then the pair $\{(g^i, \tau^i), i = L, H;\ \nu[\hat{\rho}(g,\tau), q - q^O]\}$ describes a sequential equilibrium if: (a) Voters set ν according to (15); (b) the incumbent chooses (g^i, τ^i) according to (17); and (c) voters' beliefs are *Bayes-consistent*: If $(g^L, \tau^L) \neq (g^H, \tau^H)$, then $\hat{\rho}(g^L, \tau^L) = 0$ and $\hat{\rho}(g^H, \tau^H) = 1$. If $(g^L, \tau^L) = (g^H, \tau^H)$, then $\hat{\rho}(g^L, \tau^L) = \rho$. Henceforth, I will use the term "equilibrium" as an abbreviation for "sequential equilibrium."

A. *Separating Equilibria*

In a separating equilibrium $(g^L, \tau^L) \neq (g^H, \tau^H)$. Note that in any separating equilibrium a type L must be choosing his full-information fiscal policy

$$(20)\quad (g^L, \tau^L) = [g^*(\varepsilon^L), \tau^*(\varepsilon^L)],$$

since otherwise

$$Z\{g^*(\varepsilon^L), \tau^*(\varepsilon^L), \hat{\rho}[g^*(\varepsilon^L), \tau^*(\varepsilon^L)], \varepsilon^L\} - Z(g^L, \tau^L, 0, \varepsilon^L) > 0,$$

which is inconsistent with the requirement that (g^L, τ^L) maximize (17). An incompetent incumbent gains nothing by choosing a level of fiscal policy which is distortionary and yet fails to prevent the public from deducing his type.

I will initially assume that voters' "off-the-equilibrium-path" beliefs are governed simply by $\hat{\rho}(g,\tau) = 0\ \forall\ (g,\tau) \neq (g^H, \tau^H)$. Given these beliefs, a type L will not benefit by mimicking a type H as long as $(g^H, \tau^H) \in \mathscr{A}$ where

$$(21)\quad \mathscr{A} \equiv \{(g,\tau) | Z(g, \tau, 1, \varepsilon^L) - Z[g^*(\varepsilon^L), \tau^*(\varepsilon^L), 0, \varepsilon^L] \leq 0\}.$$

In Figure 2, point I corresponds to $[g^*(\varepsilon^L), \tau^*(\varepsilon^L)]$, and set $\mathscr{A}$ consists of all points on or outside the dashed ellipse. A type L would be willing to choose any point *within* the dashed ellipse over point I if by doing so, he could fool the public into thinking he is a type H. (The assumption that U and V obey Inada conditions together with the assumption $V(0) = -\infty$ assure that the ellipse is contained within the boundaries $g > 0$, $y - \tau > 0$, and $k^L = \tau + \varepsilon^L - g > 0$. The set of points contained within the ellipse is necessarily convex as drawn, since W is strictly concave in τ and g).

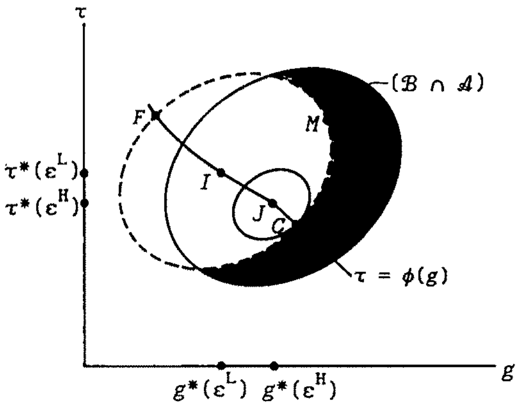


FIGURE 2. UNDOMINATED SEPARATING EQUILIBRIUM

Point J in Figure 2 corresponds to $[g^*(\varepsilon^H), \tau^*(\varepsilon^H)]$. Because all goods are normal, J must lie southeast of I. Whether J lies within the dashed ellipse (in which case it cannot be a separating equilibrium strategy for a type H) depends on a number of factors. It is more likely to be interior the larger X (ego rents), the smaller $\varepsilon^H - \varepsilon^L$, and the lower the variance of $q - q^O$. Thus if the high type is sufficiently more competent than the low type, he can choose his first-best fiscal policy and still separate himself.

Another necessary condition for a separating equilibrium is that $(g^H, \tau^H) \in \mathscr{B}$, where

$$(22)\quad \mathscr{B} \equiv \{(g,\tau) | Z(g, \tau, 1, \varepsilon^H) - Z[g^*(\varepsilon^H), \tau^*(\varepsilon^H), 0, \varepsilon^H] \geq 0\}.$$

The large solid ellipse in Figure 2 contains the convex set $\mathscr{B}$. The shaded region is $\mathscr{B} \cap \mathscr{A}$; it contains all the possible separating equilibrium strategies for a type H. It is easy to prove that $\mathscr{B} \cap \mathscr{A}$ is nonempty by virtue of the fact $\chi^H > \chi^L$ and $V'' < 0$. Thus

PROPOSITION 1: *The set of all separating equilibria is nonempty and is characterized by* $(g^L, \tau^L) = [g^*(\varepsilon^L), \tau^*(\varepsilon^L)]$, *and* $(g^H, \tau^H) \in \mathscr{B} \cap \mathscr{A}$.

B. *Undominated Separating Equilibria*

The range of separating equilibria can be drastically reduced (to a single point) by requiring that $\hat{\rho} = 1$ for all $(g, \tau) \in \mathscr{B} \cap \mathscr{A}$ and not just at (g^H, τ^H). This restriction on voters' beliefs is plausible since there are no circumstances under which a type L might benefit by choosing a point in the shaded region in Figure 2. Provided voters' beliefs reflect this minimal level of sophistication, a competent incumbent is essentially free to choose the separating strategy which is most favorable to him, that is, the one which entails the least distortions. In an *undominated* separating equilibrium, (g^H, τ^H) solves[15]

$$\max_{g,\tau} W(g, \tau, \varepsilon^H) \tag{23}$$

$$\text{s.t. } g,\ y - \tau,\ \tau + \varepsilon^H - g \geq 0,$$

and $$(g, \tau) \in \mathscr{A}.$$

PROPOSITION 2: *There exists a unique undominated separating equilibrium, and in this equilibrium* $U_1(y - \tau, g) = U_2(y - \tau, g)$.

Note that the condition $U_1 = U_2$ is precisely the same as one of the first-order conditions for a full-information optimum, equation (10). Equation (10) implicitly defines the income expansion path $\tau = \phi(g)$, which passes through points I and J in Figure 2; $\phi' < 0$ since both c and g are normal goods. The unique undominated equilibrium is given by point C in Figure 2, where $g > g^*(\varepsilon^H)$ and $\tau < \tau^*(\varepsilon^H)$. This equilibrium has the property that signaling is "efficient" in the sense that no reallocation of expenditures between the private and public consumption goods can yield voters higher welfare.

C. *Pooling Equilibria*

Restricting attention to undominated equilibria is not sufficient to rule out all pooling equilibria. For example, if ρ is large enough, then $(g^L, \tau^L) = (g^H, \tau^H) = [g^*(\varepsilon^H), \tau^*(\varepsilon^H)]$; $\hat{\rho}(g^H, \tau^H) = \rho$, can be an equilibrium. To rule out all pooling equilibria (in pure and mixed strategies), it is necessary to further refine the equilibrium concept. Following Cho and Kreps (1987), an equilibrium $\{(g^L, \tau^L), (g^H, \tau^H)\}$ is *unintuitive* if there exists a point $(\bar{g}, \bar{\tau})$ such that[16]

$$Z(\bar{g}, \bar{\tau}, 1, \varepsilon^H) - Z[g^H, \tau^H, \hat{\rho}(g^H, \tau^H), \varepsilon^H] > 0, \tag{24}$$

and

$$Z(\bar{g}, \bar{\tau}, 1, \varepsilon^L) - Z[g^L, \tau^L, \hat{\rho}(g^L, \tau^L), \varepsilon^L] < 0. \tag{25}$$

PROPOSITION 3: *All pooling equilibria are unintuitive.*

One can easily confirm that the unique undominated separating equilibrium is also an intuitive equilibrium (i.e., not unintuitive). Henceforth, the term "equilibrium" refers to this equilibrium.

---

[15] Formally, a point $(\bar{g}, \bar{\tau})$ is dominated for a type i if $Z[g^*(\varepsilon^i), \tau^*(\varepsilon^i), 0, \varepsilon^i] - Z(\bar{g}, \bar{\tau}, 1, \varepsilon^i) > 0$. Dominated equilibria are ruled out by requiring that $\hat{\rho} = 1$ at points dominated for L but not H. (As a minor technical point, $\mathscr{B} \cap \mathscr{A}$ includes points weakly dominated for L but not dominated for H).

[16] Condition (24) states that a type H would prefer to select $(\bar{g}, \bar{\tau})$ over (g^H, τ^H) if, by doing so, he could convince the public of his true type. Condition (25) states that a type L would prefer to select (g^L, τ^L) and elicit voters' equilibrium response $\hat{\rho}(g^L, \tau^L)$, than to choose $(\bar{g}, \bar{\tau})$ even if $\hat{\rho}(\bar{g}, \bar{\tau}) = 1$.

V. Multiple Elections

The extension to the case of finitely many election periods is straightforward. In off-election years, the incumbent always follows his full-information fiscal policy. Voters are able to monitor the government perfectly with a one-period lag, so there is no incentive to distort in an off-election year. During election years, the signaling problem is the same as above, except that the gain to the incumbent of reelection, χ, becomes larger as the incumbent's expected term in office increases. The prospect of being able to run for reelection again in the future raises the temptation to distort fiscal policy, and thus tends to exacerbate the political budget cycle.

The equilibrium studied here remains an equilibrium when the time horizon is infinite. It is also possible, however, to have "reputational" equilibria in which there is little or no political budget cycle if (a) the leader's rate of time preference is close to one, (b) exogenous uncertainty (here the variance of q) is not too large, and (c) the time between elections is short.[17] In most countries, though, elections are typically spaced many years apart and there is considerable uncertainty over what factors will govern distant elections. Thus incumbents are not likely to place great weight on maintaining a reputation for not engaging in political budget cycles.[18]

VI. Alternative Approaches to Mitigating the Political Budget Cycle

If preelection signaling is truly a central cause of the political budget cycle, is there any way for society to mitigate the problem?

A. *A Constitutional Amendment to Restrain Election-Year Fiscal Policy*

One natural alternative is to reform the budget process so that the government cannot alter its fiscal policy rule just prior to an election. Edward Tufte (1978, p. 152), for example, suggests instituting a change in the timing of the congressional budget cycle. Let us consider a constitutional amendment which forces the government to set fiscal policy on a biennial basis, so that both (g_t, τ_t) and (g_{t-1}, τ_{t-1}) must be set in off-election year $t-1$. (The only relevant source of uncertainty here is the α shock but in interpreting the analysis below, one can think of the government as being allowed to index its fiscal policy rule to any publicly observable shock.) If forced to bind himself in $t-1$ to (g_t, τ_t), the incumbent would solve

$$(26)\quad \max_{g,\tau} E_{t-1}[W(g,\tau,\varepsilon_t)] = \rho W(g,\tau,\alpha_{t-1}+\alpha^H) + (1-\rho)W(g,\tau,\alpha_{t-1}+\alpha^L).$$

Note that the incumbent is prevented from signaling because he does not have any information about his postelection competency type when setting election-year fiscal policy. The budget process reform thus mitigates the political budget cycle, but there are two costs. First, the public no longer has any way of distinguishing between H and L types when voting. If for simplicity we ignore the q shock, the mean cost of this lost information is $\rho\beta(\Omega^H - \Omega^O)$. The other cost is that the leader cannot employ his private information on α_t in setting period t fiscal policy.

The relative costs and benefits of the constitutional amendment are transparent in two extreme cases. If ego rents (X) are small, then the election-year fiscal policy distortions will be minor and the proposed budget process reform makes little sense. At the opposite extreme, as X becomes large, the political budget cycle takes on catastrophic dimensions. Indeed, the public may actually enjoy higher welfare during the election year itself when the incumbent is an incompetent type, since it is the competent type who distorts fiscal policy to signal his type. Nevertheless, a competent type is more likely to win reelection, since citizens vote for the candidate who offers them higher expected future welfare.

---

[17] Alesina (1987) has analyzed how reputation effects can mitigate partisan political business cycles, and John Ferejohn (1986) has considered how they can provide officeholders with incentives for taking into a account the wishes of the electorate; see also Rogoff and Sibert (1988), and Gregory D. Hess (1988).

[18] A caveat is that the government's reputation for engaging in political budget cycles may be intertwined with its general reputation for conducting a stable macroeconomic policy; see Rogoff (1989) for a survey of reputational models of macroeconomic policy.

The preceding analysis, if anything, overstates the efficacy of trying to legislate away the political budget cycle. In practice, an incumbent has a wide array of fiscal actions with which he can signal, and it is not realistically possible to constrain him in all dimensions. If this is the case, then attempts to block signaling in one set of fiscal policy instruments will tend to exacerbate distortion in others. Indeed, attempts to suppress the political budget cycle may actually reduce the welfare of the representative citizen by inducing competent types to signal inefficiently.

As a simple example, consider a "balanced budget" constraint of the form

$$\tau = \Psi(g), \qquad \Psi' > 0. \tag{27}$$

Let us initially assume that $\Psi(g)$ passes through point I in Figure 2 and also through point M. Since $\Psi(g)$ passes through $[g^*(\varepsilon^L), \tau^*(\varepsilon^L)]$, this point remains the separating equilibrium choice of a type L. The undominated separating equilibrium strategy for a type H is now found by maximizing (23) subject to the *additional* constraint (27). At the solution point, labeled M in Figure 2, the welfare of the representative voter is unambiguously lower than at C, the equilibrium in the absence of the constraint.[19]

Suppose now that $\Psi(g)$ does not pass through point I, so that the constraint (27) also distorts the choice of a type L. Denote a type L's solution to his full-information problem (9), subject to the additional constraint (27), as $[g^{**}(\varepsilon^L), \tau^{**}(\varepsilon^L)]$, and define the set $\mathcal{A}^A$ as

$$\mathcal{A}^A \equiv \{(g,\tau) | Z(g,\tau,1,\varepsilon^L) - Z[g^{**}(\varepsilon^L),\tau^{**}(\varepsilon^L),0,\varepsilon^L] \le 0\}. \tag{28}$$

The separating equilibrium strategy for a type H is again found by maximizing (23) subject to the additional constraint (27), but now with the constraint $(g,\tau) \in \mathcal{A}^A$ in place of $(g,\tau) \in \mathcal{A}$. Note that $\mathcal{A}^A$ is a subset of $\mathcal{A}$; a type L is willing to distort further to convince the public he is a type H when his best alternative is the constitutionally constrained $[g^{**}(\varepsilon^L), \tau^{**}(\varepsilon^L)]$ than when it is the unconstrained $[g^*(\varepsilon^L), \tau^*(\varepsilon^L)]$. Thus the separating strategy for a type H will generally be at some point where voter welfare is even lower than at point M.

To summarize, if a budget process reform forces a pooling equilibrium, voters may or may not be better off, depending on the costs of the lost information versus the benefits of mitigating the political budget cycle. If, however, the incumbent can still find a way to use fiscal policy to signal, the budget process reform is likely to prove counterproductive. One interesting implication of this analysis is that having an independent central bank does not necessarily reduce the *welfare* costs of political budget cycles, even if it insulates monetary policy from election-year pressures.

B. *Endogenous Elections*

Until now, I have treated the timing of elections as immutable. In principle, the framework developed here can be extended to compare social welfare under different electoral structures. For example, in a regime with very short intervals between elections, incompetent leaders can be quickly removed from office. The drawback, of course, is that the political budget cycle will occur at correspondingly high frequencies (though it may be damped since fewer years of ego rents will be at stake in each election).

---

[19] For some specifications of $\Psi(g)$ there may not exist any separating equilibrium. Of course, no pure strategy pooling equilibrium can yield higher social welfare than the solution to (26).

An especially important alternative class of electoral structures involves giving the incumbent the *option* of calling for an early election. Indeed, many modern industrialized countries, including Canada, the United Kingdom, and Japan, have systems in which the timing of elections is endogenous. In practice, early elections are sometimes forced upon a ruling party after it loses its working majority in parliament. In some instances, however, an incumbent government opportunistically decides to risk its remaining time in office in hopes of winning a fresh new term. For opportunistic early elections, the model suggests that preelection fiscal policy distortions are likely to be less severe than for end-of-term elections. The basic reason is that the call for an early election can serve as an additional (non-distorting) signal.[20]

Recall from Figure 2 that the dashed ellipse bounds the fiscal policy distortions a type L would be willing to undertake to fool the public into thinking he is a type H. Of course, this curve is drawn under the assumption that a type L's best alternative is to choose his full information policy (point I), and hope that a very favorable looks shock carries him through the election. If a type L has the superior option of waiting a period before standing for reelection, then the option of calling for an early election and mimicking the fiscal policy of a type H becomes less attractive. By waiting to call an election, a type L not only gets to enjoy a certain extra period of ego rents, but also buys time to wait for a more favorable competency shock. As a result, a competent incumbent does not need to distort fiscal policy as much to separate himself when calling an early election. Indeed, he may not need to distort at all. To assess the overall welfare implications of allowing for early elections, one must trade off the benefits of having smaller distortions with the costs of having more frequent elections.

Most cross-country empirical studies of political business cycles do not systematically distinguish between opportunistic early elections and end-of-term elections, but the above discussion suggests that the distinction may be important.[21]

C. *Self-Denial as a Signal of Competency*

Thus far I have been implicitly restricting the incumbent to signals which adversely affect all citizens and not just the incumbent himself. It is reasonable to ask whether the analysis thereby exaggerates the extent to which the incumbent will use socially destructive fiscal policy distortions to signal his competency.

Suppose, for example, that the incumbent can also signal by publicly destroying σ units of his own personal endowment of the private consumption good.[22] In this case, the incumbent's consumption is given by $c^I = y - \tau - \sigma$, and W in his objective function becomes

$$W^I(g, \tau, \sigma, \varepsilon) = U(y - \tau - \sigma, g) + \beta V(\tau + \varepsilon - g), \tag{29}$$

where $\sigma \geq 0$. Under full information, $\sigma^* = 0$, while g^* and τ^* are governed by (10) and (11) as before. Let $\tilde{Z}(g, \tau, \sigma, \hat{\rho}, \varepsilon^i)$ be the same as Z in equation (18), except with W and $\hat{\rho}(g, \tau)$ replaced by W^I and $\hat{\rho}(g, \tau, \sigma)$.

It is straightforward to show that there exist equilibria in which a competent incumbent, by setting $\sigma^H > 0$, is able to separate himself without distorting fiscal policy as much as in any of the separating equilibria where $\sigma^H = 0$. In the (unique) *undominated* separating equilibrium, however, a type H sets (g^H, τ^H, σ^H) to solve

---

[20] For an interesting formal development of the case of endogenous elections, see Marco Terrones (1989a). Terrones shows that in the unique undominated equilibrium, the budget distortions accompanying early elections are damped.

[21] Two important exceptions are Takatoshi Ito and Jin Hyuk Park (1988), and Terrones (1989b). In their study of Japan, Ito and Park find that the event of an early election does not seem to significantly impact monetary and fiscal policy. Instead, their results suggest that an incumbent government is more likely to call for an early election when recent growth and inflation performance have been strong.

[22] Milgrom and Roberts (1986) and Bagwell and Ramey (1988) model advertising by firms in a related fashion. In Bagwell and Ramey's setup, advertising is used in an undominated equilibrium only if it would have a direct positive effect on demand under full information.

$$(30)\qquad \max_{g,\tau,\sigma} W^I(g,\tau,\sigma,\varepsilon^H)$$

$$\text{s.t. } g, \sigma, y-\tau-\sigma, \tau+\varepsilon^H - g \geq 0,$$

$$\text{and}\qquad (g,\tau,\sigma) \in \tilde{\mathscr{A}},$$

where

$$\tilde{\mathscr{A}} \equiv \big\{(g,\tau,\sigma) | \tilde{Z}(g,\tau,\sigma,1,\varepsilon^L)$$

$$-\tilde{Z}\big[g^*(\varepsilon^L), \tau^*(\varepsilon^L), 0, 0, \varepsilon^L\big] \leq 0\big\}.$$

PROPOSITION 4: *In an undominated separating equilibrium,* $\sigma^H = 0$, *and* (g^H, τ^H) *solves* (23).

A type H incumbent could set $\sigma > 0$ to help signal his type, but by Proposition 4 he will always prefer to do it using fiscal policy alone. It is inefficient for a competent incumbent to signal by dissipating his personal resources, because he has no comparative advantage in that dimension.

Society could *force* the leader to dissipate $\bar{\sigma} > 0$ by requiring any incumbent who wants to run for reelection to pay a fee. It is easily shown that such a scheme can be welfare improving, but not by enough to attain the full-information equilibrium. A fee tends to distort a (selfish) leader's choice of tax policy, because it gives him a different tradeoff between private and public goods expenditure than the representative voter. But the most conspicuous drawback to this approach is that it would be very difficult in practice to find a rule for setting $\bar{\sigma}$, since incumbents differ greatly in wealth and future earning power.

Of course, if there were some way to reduce the leader's rents without causing him to distort fiscal policy, this would lead to the first-best outcome. Unfortunately, massive ego rents seem an inevitable by-product of the public goods production function.

VII. Conclusions

The present analysis preserves some of the basic insights of the Keynesian political business cycle model, albeit with significant refinements. Underlying the cycle in the model (for example, William Nordhaus, 1975) is a Keynesian Phillips curve and voter myopia. By increasing money supply growth in the year prior to an election, an incumbent national leader is able to temporarily raise output and employment.[23] Voters respond positively, not recognizing that after the election inflation will rise while output and employment will return to their natural rates. This story has two conspicuous failings. First, elections are perfectly anticipated events, so any systematic accompanying rise in money growth should not have any real effects. Second, preelection macroeconomic policy is a given by the time of the election, and voters' decisions should be governed only by which candidate offers them higher expected postelection welfare.

The Keynesian theory has generated a plethora of empirical studies aimed at testing for electoral cycles in national output, unemployment, and inflation. In light of the theoretical weaknesses of the underlying model, perhaps it should not be too surprising that the results have been mixed.[24] The equilibrium political budget cycle theory suggests that it would be more promising to focus empirical research on testing for electoral

[23] It may be possible to extend the present model to generate electoral cycles in employment. If taxes distort the labor-leisure decision, then one might expect labor supply to rise during election years when tax rates are low.

[24] Bennett T. McCallum (1978), David G. Golden and James Poterba (1980), and Nathaniel Beck (1987), among others, have suggested that there is little empirical evidence of a political business cycle in U.S. inflation and unemployment. Recently, however, Kevin B. Grier (1987), and Stephen E. Haynes and Joe A. Stone (1989) have offered a very different interpretation of the data. They argue that if one does not place arbitrary restrictions on the economy's dynamic structure, then one finds significant evidence of political business cycles. Haynes and Stone note that their test cannot discriminate between the classical political business cycle theory and the equilibrium political budget cycle theory.

cycles in taxes, transfers, and government consumption spending.[25] For these variables, one can also look at data for state and local elections, instead of concentrating solely on the small number of observations available for national elections.

In addition to focusing on different variables, the present model also offers sharper prediction concerning the dynamic structures of the cycle. Here the pre- and postelection values of observable fiscal policy variables tend to be *positively* correlated (when variables are measured in deviations from pre- and postelection means). Competent incumbents, who have greater leeway to cut taxes and raise government consumption spending prior to elections, are more likely to be able to do so after elections as well. The model here also has concrete implications for the nature of political budget cycles under alternative electoral structures. For example, in countries where the incumbent has the option of calling for early elections, the budget distortions which accompany opportunistic early elections tend to be damped compared to those accompanying end-of-term elections.

Does the political budget cycle theory have any bearing on countries such as Mexico and Japan, in which a single party dominates political life? Even in dominant-party systems, the country's leaders still generally care about their party's margin of victory. Its plurality not only affects the leaders' ability to govern the populace, but also their ability to contain internal dissent within the party. (The formal model above is easily extended to the case where plurality matters). The model should also retain some relevance in situations where competing parties share power (for example, if the majority party in the legislature is Democratic and the President is Republican). Although each of the parties may care about increasing its representation in the government, individual legislators have a strong common interest in their own reelection.

APPENDIX

This Appendix provides the proofs of Propositions 2, 3, and 4.

Proof of Proposition 2. Given the Inada conditions on U and V, any solution to (23) must have $c, g, k > 0$. Thus the Kuhn-Tucker conditions reduce to

$$U_1 - \beta V_H' = \lambda (U_1 - \beta V_L'), \tag{A1}$$

$$U_2 - \beta V_H' = \lambda (U_2 - \beta V_L'), \tag{A2}$$

$$\begin{aligned} \chi^L \pi(0) + W^*(\varepsilon^L) - \chi^L \pi(1) - U(y - \tau, g) \\ - \beta V(\tau + \varepsilon^L - g) \geq 0, \\ (= 0 \text{ if } \lambda > 0), \end{aligned} \tag{A3}$$

where $V_i' \equiv V'(\tau + \varepsilon^i - g)$. (A1) and (A2) imply that $U_1 = U_2$. This equation governs the downward-sloping income extension path $\tau = \phi(g)$ in Figure 2. (A3) is the constraint $(g, \tau) \in \mathscr{A}$, the set of points on or outside the dashed ellipse. Assume that $\lambda > 0$ so that (A3) is binding. Equations (A1)–(A3) are then satisfied at exactly two points. At point C in Figure 2, $\beta V_i' > U_1$, $i = H, L$, and $\lambda = (U_1 - \beta V_H')/(U_1 - \beta V_L') < 1$. At point F, $\beta V_i' < U_1$ and $\lambda > 1$. One can show that the second-order conditions hold if

$$(1 - \lambda)[2U_{12} - U_{11} - U_{22}](U_1 - \beta V_L')^2 > 0. \tag{A4}$$

Since all goods are normal, $2U_{12} - U_{11} - U_{22} > 0$. Thus (A4) holds only at point C.

Proof of Proposition 3. Suppose (g^z, τ^z) is any point selected with positive probability by both types. Let

$$\begin{aligned} R^i(g, \tau) \equiv Z(g, \tau, 1, \varepsilon^i) \\ - Z[g^z, \tau^z, \hat{\rho}(g^z, \tau^z), \varepsilon^i], \\ i = L, H. \end{aligned}$$

Select the pair $[\bar{g}, \phi(\bar{g})]$ such that (a) $\phi(\bar{g}) - \bar{g} < \tau^*(\varepsilon^H) - g^*(\varepsilon^H)$, and (b) $R^H[\bar{g}, \phi(\bar{g})] = 0$. Given that $V(0) = -\infty$ and $\pi(1) > \pi(\hat{\rho})$, such a pair exists and is feasible. Note that $\phi(\bar{g}) - \bar{g} < \tau^z - g^z$ since $U[y - \phi(g), g] \geq U(y - \tau^z, g^z)$ if $\phi(g) - g = \tau^z - g^z$. Then since $V'' < 0$, it follows that $R^L[\bar{g}, \phi(\bar{g})] < 0$. Thus by the continuity of R^i, $\exists\ \delta > 0$ such that $R^H[\bar{g} - \delta, \phi(\bar{g} - \delta)] > 0$ and $R^L[\bar{g} - \delta, \phi(\bar{g} - \delta)] < 0$.

---

[25] Tufte (1978) finds that in the United States, transfers rise significantly prior to presidential elections; see also Alesina (1988) and Eric Ghysels (1988). In their study of U.S. federal taxes for the years 1879–1986, Bizer and Durlouf (1989) conclude that taxes are typically reduced two years prior to successful presidential reelection attempts. Using data from twelve industrialized countries., Alesina (1989) presents evidence that federal government budget deficits tend to rise prior to elections.

The geometric intuition is that there must always exist some point on $\tau = \phi(g)$ sufficiently far southeast of J in Figure 2 such that both (24) and (25) hold. (Note in the absence of the q shock, $\pi(1)$ is not necessarily greater than $\pi(\hat{p})$, and the intuitive criterion is not generally sufficient to rule out all pooling equilibria).

Proof of Proposition 4. Any solution to (30) satisfies Kuhn-Tucker conditions analogous to (A1)–(A3), plus the additional conditions

$$\text{(A5)} \qquad (\lambda - 1)U_1 + \mu \le 0 \quad (= 0 \text{ if } \sigma > 0),$$

$$\text{(A6)} \qquad \sigma \ge 0 \quad (= 0 \text{ if } \mu > 0).$$

Assume in contradiction to the proposition that $\sigma > 0$. Then $\mu = 0$, (A5) must hold with equality and hence $\lambda = 1$. But then (A1) and (A2) require that

$$V'(\tau + \varepsilon^H - g) = V'(\tau + \varepsilon^L - g),$$

which is impossible. Thus $\sigma = 0$ and the solution to (30) is the same as to (23).

REFERENCES

Alesina, Alberto, "Macroeconomic Policy in a Two-Party System as a Repeated Game," *Quarterly Journal of Economics*, August 1987, *102*, 651–78.

______, "Macroeconomics and Politics," in Stanley Fischer, ed., *NBER Macroeconomics Annual*, Cambridge: MIT Press, 1988, 13–61.

______, "Politics and Business Cycles in Industrial Democracies," *Economic Policy*, April 1989, *8*, 54–87.

Bagwell, Kyle and Ramey, Garey, "Advertising and Limit Pricing," *Rand Journal of Economics*, Spring 1988, *19*, 59–71.

Beck, Nathaniel, "Elections and the Fed: Is There a Political Monetary Cycle?," *American Journal of Political Science*, February 1987, *31*, 194–216.

Bizer, David S. and Durlouf, Steven N., "The Behavior of U.S. Tax Rates: 1879–1986," unpublished manuscript, Stanford University, July 1989.

Chappell, Henry and Keech, William, "Welfare Consequences of a Six-Year Presidential Term Evaluated in the Context of a Model of the U.S. Economy," *American Political Science Review*, March 1983, *77*, 75–91.

Cho, In-Koo and Kreps, David M., "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, May 1987, *102*, 179–221.

Cohen, Daniel, "What Caused the Rise of Conservatism: A French View," *Economic Policy*, April 1988, *6*, 195–212.

Cukierman, Alex and Meltzer, Allan H., "A Positive Theory of Discretionary Policy, the Costs of Democratic Goverment, and the Benefits of a Constitution," *Economic Inquiry*, July 1986, *24*, 367–88.

Fair, Ray C., "The Effect of Economic Events on Votes for President," *Review of Economics and Statistics*, May 1978, *60*, 159–73.

Ferejohn, John A., "Incumbent Performance and Electoral Control," *Public Choice*, No. 1, 1986, *50*, 5–25.

Ghysels, Eric, "The Political Economy of the Budget: Some Empirical Tests," unpublished manuscript, University of Montreal, November 1988.

Golden, David G. and Poterba, James M., "The Price of Popularity: The Political Business Cycle Re-examined," *American Journal of Political Science*, November 1980, *24*, 696–714.

Grier, Kevin B., "Presidential Elections and Federal Reserve Policy: An Empirical Test," *Southern Economic Journal*, October 1987, *54*, 475–86.

Haynes, Stephen E. and Stone, Joe A., "An Integrated Test for Electoral Cycles in the U.S. Economy," *Review of Economics and Statistics*, *71*, forthcoming August 1989.

Hess, Gregory D., "Voting and the Issuance of Debt in a Macro Economy," unpublished manuscript, Johns Hopkins University, October 1988.

Hibbs, Douglas A., Jr., "Political Parties and Macreconomic Policy," *American Political Science Review*, December 1977, *71*, 146–87.

Ito, Takatoshi and Park, Jin Hyuk, "Political Business Cycles in the Parliamentary System," *Economic Letters* 3:1988, *27*, 233–38.

Keech, William and Simon, Carl, "Electoral and Welfare Consequences of Political Manipulation of the Economy," *Journal of Economic Behavior and Organization*, June 1985, *6*, 177–202.

Lindbeck, Assar, "Stabilization Policy in Open Economies with Endogenous Politicians," *American Economic Review*, May 1976, *66*, 1–19.

McCallum, Bennett T., "The Political Business Cycle: An Empirical Test," *Southern Economic Journal*, January 1978, *44*, 504–15.

Milgrom, Paul and Roberts, John, "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, August 1986, *94*, 796–821.

Moulin, Hervé, *Game Theory for Social Sciences*, New York: New York University Press, 1981.

Nordhaus, William D., "The Political Business Cycle," *Review of Economic Studies*, April 1975, *42*, 169–90.

Peltzman, Sam, "Economic Conditions and Gubernatorial Elections," *American Economic Review*, May 1987, *77*, 293–97.

Rogoff, Kenneth, "Reputation, Coordination and Monetary Policy," in Robert J. Barro, ed., *Handbook of Modern Business Cycle Theory*, Cambridge: Harvard University Press, 1989.

______ **and Sibert, Anne,** "Elections and Macroeconomic Policy Cycles," *Review of Economic Studies*, January 1988, *55*, 1–16.

Sheffrin, Steven M., "Two Tests of Rational Partisan Business Cycle Theory," unpublished manuscript, University of California at Davis, November 1988.

Terrones, Marco, (1989a) "Macroeconomic Policy Cycles Under Alternative Electoral Structures," unpublished manuscript, University of Wisconsin, February 1989.

______, (1989b) "Economic Conditions and the Timing of Elections," unpublished manuscript, University of Wisconsin, February 1989.

Tufte, Edward R., *Political Control of the Economy*, Princeton: Princeton University Press, 1978.

Willet, Thomas D., ed., *Political Business Cycles: The Political Economy of Money, Unemployment and Inflation*, Durham: Duke University Press, 1989.

Voting on the Budget Deficit

By GUIDO TABELLINI AND ALBERTO ALESINA*

This paper analyzes a model in which a group of rational individuals votes over the composition and time profile of public spending. All voters agree that a balanced budget is ex ante optimal. However, if there is disagreement between current and future majorities, a balanced budget is not a political equilibrium under majority rule. Under certain conditions a majority of the voters favors a budget deficit, and the equilibrium deficit is larger the greater is the polarization among voters. (JEL 320,025)

Opinion polls show that American voters disapprove of the federal budget deficit. However, it is politically difficult to reach an agreement on how to balance the budget: several polls show that even though voters dislike deficits, they are not in favor of any specific measure to reduce them.[1]

Two explanations for this apparent inconsistency of opinions are commonly proposed. One is that voters do not understand the concept of budget constraint, and suffer from "fiscal illusion." However, this notion is difficult to reconcile with standard assumptions of rationality.[2] The other is that disagreement generates cycling and prevents the existence of a stable majority in favor of balancing the budget. As a result, individual preferences about intertemporal fiscal policy cannot be aggregated, and no action can be taken to balance the budget. However, this argument is consistent with any outcome (deficit, surplus, balance) since the political equilibrium is indeterminate.

This paper provides an alternative explanation of budget deficits, that is based upon the inability of current voters to bind the choices of future voters. This lack of commitment, coupled with disagreement between current and future majorities, introduces a time inconsistency in the dynamic social choice problem that determines the size of budget deficits or surpluses. The policies desired by the current majority would not be carried out if future majorities exhibit different preferences. This induces the current majority to choose a debt policy that is not *ex ante* optimal for society as a whole. The deviation from optimality can be in the direction of excessive surpluses or deficits. The paper shows that a large class of individual utility functions leads to a social choice of budget *deficits*. This explains why it is hard to agree on how to eliminate deficits, even if there is a consensus that they may be socially suboptimal.

Our results have a simple economic intuition. Consider a rational voter who is presented with a number of options on how much to spend in the current period, and over what items. He votes not only on the intertemporal profile of spending, but also

---

*University of California, Los Angeles, Center for Economic Policy Research and National Bureau of Economic Research, Harvard University, Center for Economic Policy Research and National Bureau of Economic Research, respectively. We would like to thank three anonymous referees, Alex Cukierman, Patrick Kehoe, Thomas Romer, Jorgen Weibull, and participants in workshops at the Bank of Italy, Brandeis, CalTech, UC Davis, Harvard, the Hoover Institution, the Institute for International Economic Studies (Stockholm), the Bank of Israel, Tel Aviv University, University of Pittsburgh, Princeton, and UC Santa Barbara for several helpful comments.

[1] Both recent polls (*New York Times*, November 1987) and polls taken in the early 1980s (Allan Blinder and Douglas Holtz-Eakin, 1983) show that a large majority of Americans is in favor of balanced budget amendments. A much lower fraction of voters is in favor of any specific measure to reduce budget deficits.

[2] For recent arguments explaining the deficit as the result of "fiscal illusion," see James Buchanan et al. (1987) and the references quoted therein. Kenneth Rogoff and Anne Sibert (1989) show that suboptimal budget deficits may be observed if voters are rational but imperfectly informed, but only before elections and not over long time periods.

on how to allocate the resources acquired by issuing debt (or lost through a surplus). Suppose that there is uncertainty about the future composition of public spending, because the identity of future majorities is still unknown. Then, whereas the majority who runs a budget deficit also chooses how to allocate the debt proceeds, the allocation of the burden of repaying the debt is not under its control. Under appropriate conditions this asymmetry prevents the current majority from fully internalizing the costs of budget deficits, the more so the greater is the difference between its preferences and the expected preferences of the future majority.

The paper also shows that if this asymmetry is removed, and the vote on the deficit is taken behind a "veil of ignorance" on how the debt proceeds are spent, then the voters unanimously choose a balanced budget. That is, in this model a balanced budget is *ex ante* efficient. This implies that current voters would like to precommit future governments to a balanced budget rule, but no majority wants to be bound by the rule. Thus, a balanced-budget rule is enforceable only if a qualified majority is required to abrogate it.

Our results are related to those of other papers on intertemporal politico-economic models of fiscal policy. In particular, Alberto Alesina and Guido Tabellini (1987, 1989) and Guido Tabellini (1989) analyze a general equilibrium model in which two ideologically motivated parties randomly alternate in office and disagree on the optimal composition of public spending, or on the level of taxation of different constituencies. Torsten Persson and Lars Svensson (1989) considers a government that knows that its successor will want to increase public spending. In these papers as in ours, public debt is a strategic variable that affects the actions of future policymakers.

In this earlier literature, however, either the political equilibrium is exogenously given (as in Torsten Persson and Lars Svensson, 1989), or voters have to choose between two ideological candidates with fixed positions (as in Alberto Alesina and Guido Tabellini, 1987, 1989 and Guido Tabellini, 1989). In the latter case, in equilibrium both parties choose the same deficit, even though they choose a different composition of government spending. Thus, in effect, in these papers voters do not have a choice on the deficit. In particular, the question remains of whether the deficit would disappear if a centrist party promising to balance the budget enters the political arena. In the present paper there are no constraints on the policy options available to the voters. Any proposal can be voted upon in a pairwise comparison, and the voters directly vote on the size of the deficit.

The idea that state variables can be used to influence future voting outcomes is applicable to other public choice problems, besides those concerning budget deficits. For example, Ami Glazer (1987) exploits this insight to investigate the choice of durability in public capital projects. He shows that uncertainty about future voting outcomes generates a bias toward overinvesting in long run projects. Other possible applications are to privatization decisions and defense policy.

Finally, our argument is completely different from the idea that deficits occur because the current generation does not internalize the costs of taxing future generations: in our model everybody has the same time horizon. In an overlapping generations model with no altruism, on the other hand, current voters would be unanimously in favor of the largest possible budget deficit, so as to redistribute the income of future generations toward themselves. In such a model, the equilibrium would always be a corner solution and the size of budget deficits would be determined exclusively by the borrowing capacity of the government.[3]

The rest of the paper is organized as follows. Section I describes the model. The political equilibrium is computed in Section II. Section III discusses normative and positive implications for the issue of balance budget amendments. The last section suggests some extensions.

---

[3] Alex Cukierman and Allan Meltzer (1989) analyze an overlapping generations model in which individuals have a bequest motive of various intensities. The equilibrium budget deficit in that model reflects the preferences for intergenerational redistribution of current voters. Our approach and that of Cukierman and Meltzer are by no means contradictory, although very different.

I. The Model

A group of heterogeneous individuals decides by majority rule on the consumption of two public goods, g and f. The group is endowed with one unit of output in each period, and it can borrow or lend to the rest of the world at a given real interest rate, with no loss of generality assumed to be 0. The world lasts two periods, and all the outstanding debt has to be repaid in full at the end of the second period. Thus, the group faces the intertemporal constraint:

$$g_1 + f_1 - b \leq 1 \tag{1a}$$

$$f_2 + g_2 + b \leq 1, \tag{1b}$$

where subscripts denote time periods and b denotes debt. In addition, the nonnegativity constraints hold: $g_i, f_i \geq 0$, $i = 1,2$. Hence, (1b) immediately implies $-1 \leq b \leq 1$. Throughout the paper we assume that in equilibrium $-1 < b < 1$. The extension to the case $b = 1$ is straightforward, and just involves some changes in notation. At the beginning of each period, the group votes on how much to consume of each public good in that period. Thus, in period 1 the group cannot precommit to consume a specific quantity of g_2 and f_2 in the following period.

The preferences of the ith member of the group are:

$$W^i \equiv E\left\{ \sum_{t=1}^{2} \left[\alpha^i u(g_t) + (1-\alpha^i)u(f_t)\right]\right\} \tag{2}$$

where $u(\cdot)$ is concave, strictly increasing, twice continuously differentiable, and satisfies the Inada condition: $u'(0) \rightarrow \infty$. $E(\cdot)$ denotes the expectation operator. With no loss of generality, we assume that voters do not discount the future; thus the rate of time preference is equal to the world real interest rate. This eliminates any incentive to borrow or lend other than those which are the explicit focus of this paper.

The parameter α^i which identifies voter i is distributed over the $[0,1]$ interval. With only a minor change in notation, all the results can be extended to allow for values of α^i greater than 1 or negative.

This specification of individual preferences allows for disagreement about which proportion of the two public goods to consume. However, it implies that all individual preferences belong to the class of "intermediate preferences" defined by Jean Michel Grandmont (1978).[4] This class has the following useful property: individual preferences are indexed by the parameter α^i and the distribution of preferences within the group is fully summarized by the distribution of α^i. As shown by Grandmont (1978), since α^i is a scalar, preferences are single peaked and the median voter result applies: provided that all policy options are compared pairwise, the group decisions under majority rule coincide with the most preferred policy of the individual corresponding to the median value of α, denoted α^m. Thus, the political equilibrium can be computed by solving the problem of maximizing (2) subject to (1), with $\alpha^i = \alpha^m$ in (2).

A crucial feature of the model is that even though individual preferences remain stable over time, the identity of the median voter need not be the same in periods 1 and 2 (this is the reason for having the expectations operator in equation (2)). Changes in the identity of the median voter over time may be due to: (i) random shocks to the costs of voting that affect the participation rate (see John Ledyard, 1984, for a formalization of this idea); or (ii) changes in the eligibility of the voting population (for instance, because of minimum age requirements, or because of geographical movements of the population). As discussed in subsection II.D below, the extent to which these events change the median voter's preferences, in turn, depends on the underlying distribution of individual preferences.

---

[4] Any expected utility function that is linear in a vector of parameters belongs to this class. Linearity is not essential in Grandmont (1978), but it is here, since we consider an *expected* utility function. The essential property of intermediate preferences is that supporters of distinct proposals are divided by a hyperplane in the space of most preferred points. See also Andrew Caplin and Barry Nalebuff (1988).

This simple setup can be interpreted as a stylized version of several richer models. The most direct interpretation is that of a "club" with a fixed endowment to be allocated to different uses. With minor changes, the club can be interpreted as a country in which taxes are fixed and economic agents have access to a linear storage technology or to international capital markets. In an interior equilibrium, the real rate of interest on public debt equals the technologically given rate of return on storage or the world rate of interest. The extension to a model with endogenous distortionary taxation significantly increases the complexity of the analysis, without qualitatively changing the basic message of this paper. Alberto Alesina and Guido Tabellini (1987) illustrate this point in a model with a much simpler political structure.

II. Political Equilibrium

A. *The Last Period*

Consider the last period, and let α_2^m denote the value of α^i corresponding to the median voter in period 2. Two cases are possible, depending on the value of α_2^m.

If $1 > \alpha_2^m > 0$, then the median voter is at an interior optimum. In this case, his choices satisfy the following first-order condition:

$$(3)\quad \alpha_2^m u'(g_2) - (1-\alpha_2^m) u'(1-b-g_2) = 0.$$

Equations (3) and (1b) implicitly define the equilibrium values g_2^* and f_2^* as a function of α_2^m and b. Let us indicate these functions as $g_2^* = G(\alpha_2^m, b)$ and $f_2^* = 1 - b - g_2^* \equiv F(\alpha_2^m, b)$. The implicit function theorem applied to (3) and (1b), shows that, for $1 > \alpha_2^m > 0$, $G_\alpha = -F_\alpha > 0$, $-1 < G_b < 0$ and $-1 < F_b < 0$, where G_α, G_b, F_α, and F_b denote the partial derivative of $G(\cdot)$ and $F(\cdot)$ with respect to a_2^m and b, respectively.

If, instead, $\alpha_2^m = 1$ or $\alpha_2^m = 0$, then the median voter of period 2 is at a corner. If $\alpha_2^m = 1$, he sets $g_2^* = 1 - b$ and $f_2^* = 0$; thus $G_b = -1$ and $F_b = 0$. Symmetric results hold if $\alpha_2^m = 0$.

B. *The First Period: Preliminary Results*

In period 1 there is uncertainty about the identity of the median voter of period 2. Hence, from the point of view of the voters in period 1, the parameter α_2^m in (3) is a random variable. The policy most preferred by the median voter of period 1 (whose preferences are denoted by α_1^m) can be found by solving the following optimization problem:

$$(4)\quad \max_{g_1, b} \Big\{ \alpha_1^m u(g_1) + (1-\alpha_1^m) u(1-g_1+b) + E\big[\alpha_1^m u(G(\alpha_2^m, b)) + (1-\alpha_1^m) u(F(\alpha_2^m, b))\big] \Big\}.$$

The current median voter maximizes an expected utility function, since in the second period g_2 and f_2 may be chosen by a different majority. The expectation operator is taken with respect to α_2^m. Thus, today's voters choose the value of the state variable b taking into account how this choice influences the policies chosen by future majorities.[5]

If $1 > \alpha_1^m > 0$, the first-order condition relative to g_1 is:

$$(5)\quad \alpha_1^m u'(g_1) - (1-\alpha_1^m) u'(1+b-g_1) = 0.$$

Equation (5) implicitly defines the optimal values g_1^* and f_1^*, as a function of α_1^m and b: $g_1^* = g(\alpha_1^m, b)$, $f_1^* = f(\alpha_1^m, b)$. Using the same notation as before, it can be shown that, for $1 > \alpha_1^m > 0$, $1 > g_b > 0$, and $f_b = 1 - g_b$. If instead $\alpha_1^m = 1$ (or $\alpha_1^m = 0$), then the median voter in period 1 is at a corner and chooses respectively $f_1^* = 0$ and $g_1^* = 1 + b$ (or $g_1^* = 0$ and $f_1^* = 1 + b$).

---

[5] This setting is reminiscent of that analyzed in Robert Strotz (1956) and Bezalel Peleg and Menahem Yaari (1973), where a consumer with time inconsistent preferences solves a dynamic optimization problem. In those papers, like here, the time consistent solution is the noncooperative equilibrium of a game played by successive decision makers.

The intertemporal choice is described by the first-order condition of problem (4) relative to b, which for $b < 1$ is:

$$\begin{aligned}(6)\quad & \alpha_1^m u'(g(\alpha_1^m, b)) \\ & + E\big[\alpha_1^m u'(G(\alpha_2^m, b))G_b \\ & \quad + (1 - \alpha_1^m) \\ & \quad \times u'(F(\alpha_2^m, b))F_b\big] = 0,\end{aligned}$$

where G_b and F_b are functions of α_2^m and b. Despite the concavity of $u(\cdot)$, the second-order conditions are not satisfied unless an additional mild condition is imposed. We assume throughout the paper that this condition is satisfied for any value of α_2^m and α_1^m.[6]

The first term on the left-hand side of (6) is the marginal gain of issuing one more unit of debt; at the optimum, this must coincide with the marginal utility of spending one extra unit on either of the two public goods (good g in (6)). The second term of (6) is the expected marginal disutility of repaying the debt, by cutting public spending tomorrow. This term takes into account that the future composition of public spending depends on α_2^m. The solution to equation (6) determines the equilibrium value of debt, b^*, chosen by majority rule in period 1.

In order to sign b^* in the next subsection we consider equation (6) at the point $b = 0$. If at this point equation (6) is satisfied, then $b^* = 0$. If instead at $b = 0$ the left-hand side of (6) is positive (negative), then by the second-order condition $b^* > 0$ ($b^* < 0$).

C. *The Equilibrium Level of Debt*

Consider first the case in which the median voter at time 1 is certain that he will also be the median voter in period 2 (i.e., $\alpha_1^m = \alpha_2^m$ with certainty). The second term in (6) reduces to $\alpha_1^m u'(G(\alpha_1^m, b))$, so that $b^* = 0$ is the only solution to (6) for any value of α_1^m. Intuitively, since the discount rate coincides with the real interest rate (they are both zero), in the absence of political instability the median voter chooses to spend an equal aggregate amount in both periods. It is easy to show that $b^* = 0$ is also the policy that would be chosen by a social planner maximizing a weighted sum of individual utilities, for any choice of weights in the planner's objectives. Thus, with no disagreement between current and future majorities, the political equilibrium lies on the Pareto frontier.

The remainder of this section investigates the case in which $\alpha_2^m \neq \alpha_1^m$ with positive probability. It is convenient to divide the second term on the left-hand side of (6) into the weighted average of two conditional expectations: the expectation conditional on the event that $1 > \alpha_2^m > 0$; and the expectation conditional on the event that $\alpha_2^m = 1$ or $\alpha_2^m = 0$.

Although special, the second case provides the simplest illustration of why political in-

---

[6] This second-order sufficient condition can be stated as follows:

$$\begin{aligned}(\text{F.1})\quad & R(f_2)^3 R(g_2) + R(g_2)^2 R(f_2)^2 \\ & + (1 - \gamma) R'(g_2) R(f_2)^2 \\ & + \gamma R(g_2)^3 R(f_2) \\ & + \gamma R(g_2)^2 R(f_2)^2 \\ & + (\gamma - 1) R'(f_2) R(g_2)^2 > 0,\end{aligned}$$

where

$$\gamma = \frac{1 - \alpha_1^m}{\alpha_1^m} \frac{1 - \alpha_2^m}{\alpha_2^m}$$

and where

$$R(\cdot) = -u''(\cdot)/u'(\cdot)$$

is the coefficient of absolute risk aversion of $u(\cdot)$.In turn, a sufficient (but not necessary) condition for (F.1) to hold is that $R(f_2)R(g_2) + R(g_2)^2 + R'(g_2) > 0$ and $R(f_2)R(g_2) + R(f_2)^2 + R'(f_2) > 0$.

stability creates incentives to issue public debt. In this case future median voters are expected to be at a corner, so that they spend in only one kind of public good: g_2 if $\alpha_2^m = 1$, and f_2 if $\alpha_2^m = 0$. If $\alpha_2^m \neq \alpha_1^m$ with positive probability, we have:

PROPOSITION 1: (*i*) *If either* $\alpha_2^m = 0$ *or* $\alpha_2^m = 1$, *then* $b^* > 0$. (*ii*) b^* *is greater the larger is the difference between* α_1^m *and the expected value of* α_2^m.

PROOF:

(i) Let $\alpha_2^m = 1$ with probability π and $\alpha_2^m = 0$ with probability $1-\pi$, $1 > \pi > 0$. Then, using (5), equation (6) can be rewritten as:

$$\begin{aligned}(7)\qquad \alpha_1^m u'(g_1^*) - \tilde{\alpha} u'(1-b) &= (1-\alpha_1^m) u'(f_1^*) \\ &\quad - \tilde{\alpha} u'(1-b) \\ &= 0,\end{aligned}$$

where $\tilde{\alpha} = \alpha_1^m \pi + (1-\pi)(1-\alpha_1^m)$. Clearly, $\tilde{\alpha} \leq \mathrm{Max}(\alpha_1^m, (1-\alpha_1^m))$, with strict inequality if $\alpha_1^m \neq 1/2$. Moreover, at the point $b=0$, $u'(1-b) \leq u'(g(\alpha_1^m, b))$ and $u'(1-b) \leq u'(f(\alpha_1^m, b))$, with strict inequality if $1 > \alpha_1^m > 0$. Hence, at the point $b=0$ the two terms in the left-hand sides of (7) are always strictly positive. By the second-order conditions this implies $b^* > 0$. (ii) The expected value of α_2^m is π. Fix α_1^m, and consider b^* as a function of π. We have:

$$(8)\qquad \frac{db^*}{d\pi} = \frac{db^*}{d\tilde{\alpha}} \frac{d\tilde{\alpha}}{d\pi} = \frac{db^*}{d\tilde{\alpha}} (2\alpha_1^m - 1).$$

Applying the implicit function theorem to (7), we obtain that $db^*/d\tilde{\alpha} < 0$. Hence,

$$(9)\qquad \frac{db^*}{d\pi} \lesseqgtr 0 \quad \text{as} \quad \alpha_1^m \gtreqless \frac{1}{2}.$$

Thus, if $\alpha_1^m > 1/2$, a lower value of π increases b^*. And conversely, if $\alpha_1^m < 1/2$, a higher value of π increases b^*. Hence, b^* increases with the difference between α_1^m and the expected value of α_2^m. □

The intuition is that an increase in debt today implies a reduction of aggregate spending tomorrow. But tomorrow only one kind of public good will be consumed. Hence, with positive probability (and with probability 1 if $1 > \alpha_1^m > 0$), this reduction of spending will affect only the good with a low marginal utility from the point of view of today's median voter. Thus, the median voter of period 1 does not fully internalize the cost of issuing debt: he finds it optimal to spend in excess of the current aggregate endowment. Moreover, the incentive to borrow is stronger the lower is the marginal utility of the future public good. This is more likely to happen if the future median voter has very different tastes from the current median voter.

We now show that, under appropriate conditions, this basic intuition extends to the more general case in which α_2^m lies in the open interval (0,1). With no loss of generality, suppose that over this interval α_2^m is distributed according to a continuous probability function $H(\cdot)$, where $H(\alpha) \equiv \mathrm{prob}(\alpha_2^m \leq \alpha)$. Then (6) can be rewritten as:

$$(10)\qquad \int_0^1 [\alpha_1^m u'(g_1^*) - v(\alpha_2^m)]\, dH(\alpha_2^m) = 0,$$

where $v(\alpha_2^m)$ is the marginal cost of repaying the debt, given that in period 2 the median voter tastes parameter is α_2^m. After some transformations we obtain:

$$(11)\qquad v(\alpha_2^m) = \frac{u'(g_2^*) u'(f_2^*) [\alpha_1^m \lambda(f_2^*) + (1-\alpha_1^m) \lambda(g_2^*)]}{u'(g_2^*) \lambda(g_2^*) + u'(f_2^*) \lambda(f_2^*)},$$

where $g_2^* = G(\alpha_2^m, b)$, $f_2^* = F(\alpha_2^m, b)$, and where $\lambda(\cdot) \equiv -u''(\cdot)/[u'(\cdot)]^2$ is the "concavity index" of $u(\cdot)$ as in Gerard Debreu and Tjalling C. Koopmans (1982).

We now assume that $u(\cdot)$ has the following property.[7]

---

[7] This condition can also be stated as:

$$u'''(x) > 2[u''(x)]^2/u'(x), \qquad 1 > x > 0.$$

(c) The concavity index of $u(x)$, $\lambda(x)$, is decreasing in x, for $1 > x > 0$.

That is, we assume that $u(\cdot)$ becomes less concave in the sense of the index of Gerard Debreu and Tjalling C. Koopmans (1982) as consumption increases. This hypothesis is more restrictive than decreasing absolute risk aversion: it implies that the coefficient of absolute risk aversion falls more rapidly than marginal utility as consumption increases. This hypothesis is satisfied for several commonly used utility functions, such as any CES function $u(x) = x^{\gamma}/\gamma$ with $1 > \gamma > 0$.

The Appendix proves that, at the point $b = 0$, $\alpha_1^m u'(f_1) - v(\alpha_2^m) > 0$ for any $\alpha_2^m \neq \alpha_1^m$ if $u(\cdot)$ satisfies condition (c). Hence, under this condition, at the point $b = 0$ the marginal gain of issuing debt exceeds the corresponding expected marginal cost (i.e., the left-hand side of (10) is strictly positive at the point $b = 0$). Thus:

PROPOSITION 2: *Given that* $\alpha_2^m \in (0,1)$, $b^* > 0$ *if* (*c*) *holds.*

Next, let us define the probability distribution $H(\alpha_2^m)$ as "more polarized relative to α_1^m" than the distribution $K(\alpha_2^m)$ if, for any continuous increasing function $f(\cdot)$, the following condition is satisfied:

$$(12)\quad \int_0^1 f(|\alpha_2^m - \alpha_1^m|)\, dH(\alpha_2^m) > \int_0^1 f(|\alpha_2^m - \alpha_1^m|)\, dK(\alpha_2^m).$$

That is, a more polarized probability distribution assigns more weight to values of α_2^m that are further apart from α_1^m. The Appendix also proves that, if condition (c) holds, then for any $b > 0$ the expression $[\alpha_1^m u'(g_1^*) - v(\alpha_2^m)]$ is an increasing function of $|\alpha_2^m - \alpha_1^m|$ (strictly increasing if $|\alpha_2^m - \alpha_1^m| > 0$). Then, using (10) and appealing to the second-order conditions, we also have:[8]

PROPOSITION 3: *Under the same condition of Proposition* 2, b^* *is larger the more polarized is the probability distribution of* α_2^m *relative to* α_1^m *over the interval* (0,1).

If the concavity index $\lambda(x)$ is everywhere increasing (constant) for $1 > x > 0$, then Propositions 2 and 3 hold in reverse: $b^* < 0$ ($b^* = 0$), and b^* is more negative if $H(\alpha_2^m)$ is more polarized. If $\lambda(x)$ is not monotonic over $1 > x > 0$, then the sign of b^* is ambiguous.

The role played by condition (c) is highlighted in Figure 1. The downward sloping line denotes the opportunity set faced by the median voters in both periods if $b = 0$. A positive value of b shifts this line to the right in period 1, and to the left in period 2. A and B denote the points chosen in periods 1 and 2 by the median voters of type α_1^m and α_2^m respectively, again for $b = 0$. For concreteness, it has been assumed that $\alpha_1^m > 1/2 > \alpha_2^m$. The indifference curves for the median voter of type α_1^m in periods 1 and 2 are labeled I and II, respectively. Finally, the upward sloping lines EP_1 and EP_2 denote the income expansion paths of types α_1^m and α_2^m. With a decreasing concavity index, the voters' indifference curves become flatter at higher levels of income; that is, the two public goods become closer substitutes. As a result, the divergence between the choices of the two types of median voter increases with income, and their income expansion paths diverge.[9] To put it differently, with a decreasing concavity index for $u(\cdot)$, disagreement concerning the optimal composition of g and f is a luxury good: it grows with the overall size of public spending.

The ambiguity of the sign of b^* for $1 > \alpha_2^m > 0$ is due to the opposite influence of two countervailing forces. By running a surplus

---

[8] The same results would go through if other measures of distance between α_2^m and α_1^m were used in (12), such as euclidean norm or $(\alpha_1^m - \alpha_2^m)^2$.

[9] The income expansion paths are not necessarily linear: Their slopes can be shown to equal $R(g_2^*)/R(f_2^*)$ and $R(f_2^*)/R(g_2^*)$ for EP_2 and EP_1, respectively, where $R(\cdot) \equiv -u''(\cdot)/u'(\cdot)$ is the coefficient of absolute risk aversion of $u(\cdot)$. Note that the income expansion paths would be divergent even if points A and B were both below the 45° line, that is, if either $\alpha_1^m, \alpha_2^m > 1/2$, or $\alpha_1^m, \alpha_2^m < 1/2$.

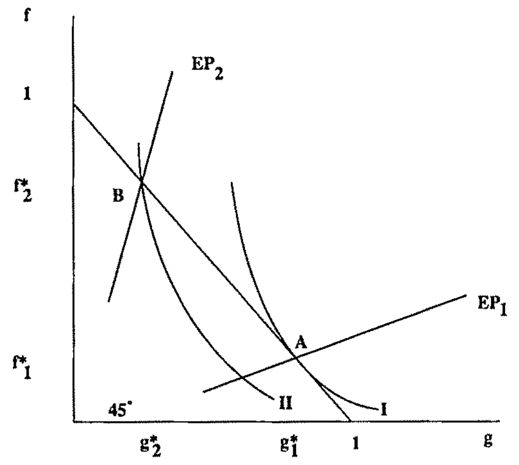


FIGURE 1

($b < 0$), the median voter in period 1 moves A to the left along EP_1 and B to the right along EP_2; this has the effect of reducing the distance between the indifference curves labeled I and II. Hence, a surplus "buys insurance" for the median voter of period 1, since it tends to equalize the median voter's utility in the two periods. This is the force that works in the direction of making a surplus today more desirable.

On the other hand by running a deficit ($b > 0$), the median voter of period 1 moves B to the left along EP_2. This takes the future composition of public spending toward the point that is preferred by today's median voter. This is the force that provides the incentive to issue public debt today. Debt is used strategically, to influence the future spending decision in the direction preferred by the current majority.

Condition (c) guarantees that the second effect dominates the first. This condition is more likely to be satisfied if the slopes of EP_1 and EP_2 are very divergent from each other (that is, if the substitutability of g and f increases very rapidly with income); or if the indifference curves are very flat (that is, if the utility function is not very concave), because in this case the indifference curves labeled I and II are close to each other.

Summarizing, Propositions 1–3 imply that an equilibrium with debt occurs if: (i) the future median voter has extreme preferences and is at a corner (i.e., $\alpha_2^m \notin (0,1)$); or (ii) the concavity index of $u(\cdot)$ is decreasing. Moreover, in both cases, the size of debt is larger the greater is the likelihood of values of α_2^m very different from α_1^m; that is, using the previous terminology, the more polarized are the current and future majorities.

In a more general model, the future median voter could be at a corner even if $1 > \alpha_2^m > 0$. For instance, if the utility function $u(\cdot)$ did not satisfy the Inada conditions, so that the indifference curve of Figure 1 would intersect either the horizontal or the vertical axis. Alternatively, if the public goods g and f had to be provided in some minimum amounts (for instance, because of survival reasons), the future decision maker could be at a corner even for $1 > \alpha_2^m > 0$. In both cases, the income expansion paths of future majorities would be either vertical or horizontal, so that issuing debt would always take the composition of public spending in the desired direction.[10]

D. *Positive Implications*

Propositions 1–3 relate the size of budget deficits to the instability of the median voters' preferences over time. This type of instability, in turn, depends upon the distribution of individual preferences within society. We now argue that the more "homogeneous" are the preferences of different individuals, *coeteris paribus* the more stable are the median voter preferences over time.

Consider a family of density functions indexed by ε: let $\gamma(\alpha, \varepsilon)$ be the frequency distribution of α over the $[0,1]$ interval, where α is the parameter that summarizes individual preferences in equation (2). Thus, ε represents a perturbation of the distribution of the voters' preferences, associated with random shocks to the voting participation or to the eligibility of the voting population.

---

[10] Note however that the probability that the future decision maker is at a corner would be endogenous in this case, and in particular it would depend on the size of the debt. This adds another dimension to the problem.

The median voter's preferences, $\alpha^m(\varepsilon)$, are then defined implicitly by:

$$\int_0^{\alpha^m}\gamma(\alpha,\varepsilon)\,d\alpha-\frac{1}{2}=0. \tag{13}$$

The relationship between α^m and ε depends on the properties of the density function $\gamma(\alpha,\varepsilon)$: by applying the implicit function theorem to (13) one obtains:

$$\frac{d\alpha^m}{d\varepsilon}=-\frac{\int_0^{\alpha^m}\gamma_\varepsilon(\alpha,\varepsilon)\,d\alpha}{\gamma(\alpha^m,\varepsilon)}, \tag{14}$$

where $\gamma_\varepsilon(\alpha,\varepsilon)\equiv\partial\gamma(\cdot)/\partial\varepsilon$. The numerator of (14) is the area underneath the density function that is shifted from one side to the other of α^m as ε varies. According to (14), for a given value of the numerator, the term $d\alpha^m/d\varepsilon$ is larger in absolute value the smaller is $\gamma(\alpha^m,\varepsilon)$. That is, if there are relatively few individuals in the population that share the median voter's preferences (i.e., if $\gamma(\alpha^m,\varepsilon)$ is small for all ε), then α^m varies a lot as the distribution is perturbed. Conversely, if the median voter preferences are representative of a large part of the population (i.e., if $\gamma(\alpha^m,\varepsilon)$ is large), then α^m is stable even in the face of large perturbations.

This result is illustrated in Figure 2. Consider the top distribution first. When ε changes from ε_1 to ε_2, a fraction of individuals corresponding to the area A is moved from the right to the left of $\alpha_1^m=\alpha^m(\varepsilon_1)$, to the area $A^1=A$. This area is the numerator of (14). The new median voter, $\alpha_2^m=\alpha^m(\varepsilon_2)$, is found by equating the area between α_1^m and α_2^m, B, to the area A. Repeat the same perturbation to the distribution in the bottom of Figure 2. Clearly, the same area B corresponds to a larger horizontal distance between α_1^m and α_2^m: since the frequency of the population around α^m is relatively small, the median voter's preferences shift by more than in the case of the upper distribution. This is the sense in which a more polarized distribution of voters' preferences is associated with more instability in the induced probability distribution of the median voter's preferences.

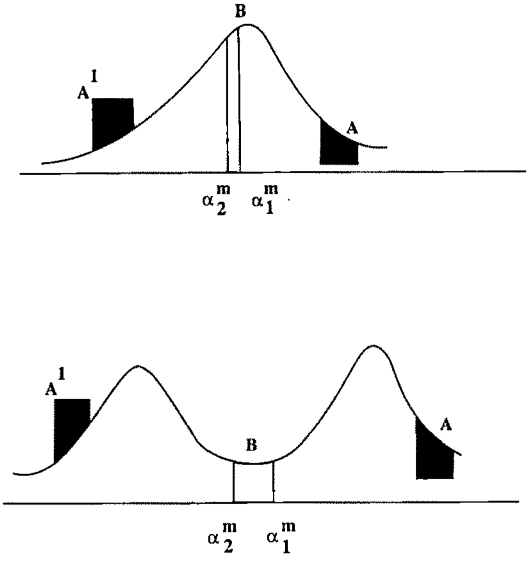


FIGURE 2

These considerations are suggestive of a testable implication that can explain the observed cross-country differences in debt policies. In more unstable and polarized political systems, there is a higher probability that future majorities will allocate government revenues to uses that are not valued by the current majority. Hence, according to our results, more polarized and politically unstable societies should exhibit larger deviations from budget balance, and if condition (c) is satisfied, these deviations should be in the direction of budget deficits.

III. Constitutional Constraints on the Budget Deficit

Section II.C shows that a social planner with stable preferences always chooses $b^*=0$, for any weighting of individual utilities. That is, a balanced budget is always a component of the first best policy. On this ground, it is tempting to conclude that a budget deficit is inefficient in this model. However, this argument should be qualified. By assumption, a social planner chooses the composition of both periods 1 and 2 public goods according to a stable social welfare function. This assumption is violated in the political equilibrium of the model and in any real world political regime: the current majority cannot precommit the spending choices of future majorities. Hence, the solution to the social planner's optimum is not necessarily the optimal social contract for a group of individuals who cannot precommit the spending choices of future governments.

In order to characterize such an optimal social contract, we need to ask what is the optimal level of debt when there is uncertainty about the median voter preferences in both period 2 *and* period 1. Suppose that b is chosen under a "veil of ignorance," before knowing the composition of public spending in period 1. Following John Rawls (1971) and James Buchanan and Gordon Tullock (1962), we can think of a constitutional amendment on budget deficits as being chosen in this way. The optimal level of b for agent i is then determined as the solution to the following problem:

$$\begin{aligned}(15)\quad \underset{b}{\operatorname{Max}}\, E\big\{\alpha^i\big[u\big(g(\alpha_1^m, b)\big) \\ + u\big(G(\alpha_2^m, b)\big)\big] \\ + (1-\alpha^i)\big[u\big(f(\alpha_1^m, b)\big) \\ + u\big(F(\alpha_2^m, b)\big)\big]\big\},\end{aligned}$$

where E is the expectations operator with respect to the random variables α_1^m and α_2^m; and where $g(\cdot)$, $f(\cdot)$, $G(\cdot)$, and $F(\cdot)$ are defined implicitly by (5) and (3) of the previous section. If α_1^m and α_2^m are drawn from the same prior distribution, then it is easy to show that the only solution to (15) is $b = 0$, for any value of α^i. Using the terminology of Bengt Holmstrom and Robert Myerson (1983), we conclude that a balanced budget rule is "*ex ante* efficient": before knowing the identity of the current majority, the group unanimously favors a balanced budget.[11]

If however the value of α_1^m is known when choosing b, then we are back in the equilibrium examined in the previous section, where a majority might support a deficit. These results may explain why the majority of voters seems to generally favor an abstract notion of balanced budgets, even though when choosing specific policies the same majority votes in favor of budget deficits (see the literature quoted in fn. 1). Balanced budgets are *ex ante* efficient; therefore, voters asked in a poll would approve of a balanced budget constitutional amendment. However, the same voters may favor a budget deficit in the current period, if uncertain about the preferences of future majorities.

More generally, each current majority does not want to be bound by the rule, even though it wants the rule for all future majorities. However, a budget rule taking effect at some prespecified future date would be irrelevant: if the rule can be abrogated by a simple majority, then any future majority would follow the policy described in Section II and would abrogate the rule. Using again the terminology of Bengt Holmstrom and Robert Myerson (1983), we conclude that in our model a balanced budget rule, though *ex ante* efficient, is not "durable" under simple majority.

This problem could be overcome by requiring a qualified majority to abrogate the rule. But this requirement would greatly reduce the flexibility with which to respond to

---

[11] If α_1^m and α_2^m have the same probability distribution, say $H(\cdot)$, then the first-order condition of (15) with respect to b can be written as:

$$\begin{aligned}\alpha^i\int_0^1\big[u'\big(g(\alpha, b)\big)g_b(\alpha, b) \\ + u'\big(G(\alpha, b)\big)G_b(\alpha, b)\big]\,dH(\alpha) \\ + (1-\alpha^i)\int_0^1\big[u'\big(f(\alpha, b)f_b(\alpha, b) \\ + u'\big(F(\alpha, b)\big)F_b(\alpha, b)\big]\,dH(\alpha) = 0.\end{aligned}$$

If $b = 0$, the terms inside each integral sum to zero. Hence, by the second-order conditions, $b = 0$ is the solution to (15). Unanimity would be lost if the distributions of α_1^m and α_2^m in (15) were different.

unexpected events. A budget rule could contain escape clauses, such as for cyclical fluctuations of tax revenues or wars. However, since it is very difficult or even impossible to list all relevant contingencies, requiring a very large majority to abandon (even temporarily) the budget balance constraint may be counterproductive. Presumably, in a model with uncertainty and constraints on the degree of "complexity" of the rule, there would be an "optimal qualified majority" corresponding to the optimal point on the tradeoff between commitment and flexibility.

Summarizing, there is a role for institutions that enable society to separate its intertemporal choices from decisions concerning the allocation of resources within any given period. Without this separation, the conflict over the allocation of resources within each period distorts society's intertemporal choices. However, there is also an inescapable conflict between preserving sufficient flexibility to meet unexpected contingencies, and the enforcement of this separation. Thus, as in many other problems of macroeconomic policy, such as monetary policy, society has to choose between simple rules and discretion.[12]

IV. Summary and Extensions

Disagreement between current and future voters about the composition of public expenditure generates a suboptimal path of public debt. Public debt is the legacy left by today's majority to the future, and under specific conditions it tends to increase with the likelihood of disagreement between current and future voters. The results of this paper are in principle testable. On time-series data for a single country, we should observe sustained budget deficits whenever a government with extreme preferences relative to the historical average wins the temporary support of a majority of the voters. On cross-countries data, more polarized and politically unstable countries should have a larger stock of debt outstanding than more homogeneous and stable societies. Nouriel Roubini and Jeffrey Sachs (1989a, b) present encouraging evidence along these lines.

Some possible generalizations of the basic framework of this paper are suggested in Section I. Another feasible extension would be to have an infinite horizon, by applying the dynamic programming solution procedure discussed in Alberto Alesina and Guido Tabellini (1987). With an infinite horizon, cooperation between current and future majorities could be sustained by trigger strategy equilibria. In these equilibria the path of public debt would approach the socially efficient value. However, this would require cooperation between successive majorities: cooperation amongst different voters within the same time period would not solve the intertemporal distortions that are the focus of this paper. Hence, this form of cooperation necessitates substantial coordination among voters. In addition, with discounting of the future, the qualitative implications of reputational equilibria are similar to those of the equilibrium studied in the present paper, as argued in a different context by Alberto Alesina (1987, 1988a).

Finally, a natural and yet difficult extension of the basic model would be to allow the voters to choose whether or not to repudiate the debt. In fact, the results of this paper are driven by an asymmetry in the possibility of commitments: even though voters cannot bind the future allocation of spending, they can force future majorities to fully service the debt. This assumption is realistic if applied to industrialized economies during recent decades. But still the puzzle remains of what is the source of this asymmetry. Some recent literature has emphasized that reputation creates incentives to honor the debt of previous governments.[13] A second answer closer to the spirit of this paper is that defaulting on the government debt brings about political and redistributional costs.[14] Further investigation of this point is the task of future research.

---

[12] Interestingly, in the case of budget deficits in the United States this conflict has been resolved in different ways at the federal and state government levels. Whereas the federal government and legislature have retained full discretion in their borrowing policies, the constitution of most states in the United States forbids the issue of state or local government debt to finance current expenditures. These state restrictions on public borrowing probably reflect the 19th century history of defaults of local and state debts (see B. V. Ratchford, 1941, and William A. Scott, 1893). But the asymmetry between the federal and state restrictions on public borrowing may also be due to the value of discretion being higher at the federal than at the state level: expenditures and revenues of state governments are generally easier to predict than those of the federal government.

APPENDIX

For a given value of b, $v(\alpha_2^m)$ is continuous in $1 > \alpha_2^m > 0$ (since $u(\cdot)$ is assumed to be twice continuously differentiable). After some algebra, $v'(\alpha_2^m)$ simplifies to:

$$v'(\alpha_2^m) = \frac{u'(g_2^*)\Delta\left(\dfrac{1-\alpha_1^m}{\alpha_1^m} - \dfrac{1-\alpha_2^m}{\alpha_2^m}\right)}{[R(g_2^*)+R(f_2^*)]^2}\,\frac{dg_2^*}{d\alpha_2^m},$$

where

$$\text{(A1)} \qquad \frac{dg_2^*}{d\alpha_2^m} > 0 \text{ and}$$

$$\Delta = R(g_2^*)\left[R(f_2^*)^2 + R'(f_2^*)\right] + \left[R(g_2^*)^2 + R'(g_2^*)\right].$$

If $\lambda(x) \equiv -u''(x)/[u'(x)]^2 \equiv R(x)/u'(x)$ is decreasing in x for $1 > x > 0$, then $\Delta < 0$. Hence for any b:

$$\text{(A2)} \qquad v'(\alpha_2^m) \gtreqless 0 \text{ as } \alpha_2^m \lesseqgtr \alpha_1^m,$$

if (c) holds. These properties imply that, under (c), $v(\alpha_2^m)$ reaches a maximum at the point $\alpha_2^m = \alpha_1^m$, and is strictly decreasing in $|\alpha_2^m - \alpha_1^m|$ if $\alpha_2^m \neq \alpha_1^m$. Hence, for given α_1^m and given b, the expression $[\alpha_1^m u'(g_1) - v(\alpha_2^m)]$ reaches a minimum at $\alpha_2^m = \alpha_1^m$ and is strictly increasing in $|\alpha_2^m - \alpha_1^m|$ if $\alpha_2^m \neq \alpha_1^m$.

Consider now this expression at the point $b = 0$. The discussion on p. 8 of the text implies that, at $b = 0$, $\alpha_1^m u'(g_1) - v(\alpha_1^m) = 0$. Since, as shown above, under (c) $\alpha_1^m = \operatorname{argmax} v(\alpha_2^m)$, we have that, if $b = 0$:

$$\alpha_1^m u'(g_1) - v(\alpha_2^m) \geq 0$$

with strict inequality if $\alpha_2^m \neq \alpha_1^m$.

REFERENCES

Alesina, Alberto, "Macroeconomic Policy in a Two-Party System as a Repeated Game," *Quarterly Journal of Economics*, August 1987, *102*, 651–78.

——, (1988a) "Credibility and Policy Convergence in a Two-Party System with Rational Voters," *American Economic Review*, September 1988, *78*, 796–805.

——, (1988b) "The End of Large Public Debts," in F. Giavazzi and L. Spaventa eds., *High Public Debt: The Italian Experience*, Cambridge University Press and CEPR, Cambridge, 1988.

—— **and Tabellini, Guido,** "A Positive Theory of Fiscal Deficits and Government Debt in a Democracy," NBER Working Paper No. 2308, July 1987.

—— **and** ——, "External Debt, Capital Flights and Political Risk," *Journal of International Economics*, 1989, forthcoming.

Barro, Robert, "On the Determination of Public Debt," *Journal of Political Economy*, Part 1, October 1979, *87*, 940–47.

Blinder, Allan, and Holtz-Eakin, Douglas, "Public Opinion and the Balanced Budget," NBER Working Paper No. 1234, 1983.

Buchanan, James, Rowley, Charles and Tollison, Robert, *Deficits*, Oxford, UK: Basil Blackwell, 1987.

—— **and Tullock, Gordon,** *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1962.

Bulow, Jeremy and Rogoff, Kenneth, "A Recontracting Model of External Debt," *Journal of Political Economy*, February 1989, *79*, 155–78.

Caplin, Andrew and Nalebuff, Barry, "On 64% Majority Rule," *Econometrica*, July 1988, *56*, 787–814.

Cukierman, Alex and Meltzer, Allan, "A Political Theory of Government Debt and Deficits in a Neo-Ricardian Framework," *American Economic Review*, September 1989, *79*, 713–31.

---

[13] See in particular Herschel Grossman and John Van Huyck (1987). A larger literature has investigated the problem of external debt repudiation, for instance, Jeffrey Sachs (1984), Jeremy Bulow and Kenneth Rogoff (1989), Herschel Grossman and John Van Huyck (1988).

[14] Recent accounts of historical episodes of debt repayments in Europe during the interwar period lend support to this second view (see for instance Alberto Alesina (1988b) and B. Eichengreen (1989)). Guido Tabellini (1989) analyzes a model in which in equilibrium a majority of the voters is in favor of repaying the public debt outstanding, so as to avoid wealth redistributions.

Debreu, Gerard and Koopmans, Tjalling C., "Additively Decomposed Quasi-convex Functions," *Mathematical Programming*, *24*, February 1982, 1–38.

Eichengreen, B., "The Capital Levy in Theory and Practice," in R. Dornbusch and M. Draghi, eds., Cambridge: Cambridge University Press, forthcoming 1989.

Ferejohn, John, "Incumbent Performance and Electoral Control," *Public Choice*, January 1986, *50*, 5–26.

Glazer, Ami, "Politics and the Choice of Durability," mimeo., University of California at Irvine, 1987.

Grandmont, Jean Michel, "Intermediate Preferences and the Majority Rule," *Econometrica*, March 1978, *46*, 317–30.

Grossman, Herschel and Van Huyck, John, "Nominally Denominated Sovereign Debt, Risk Shifting and Reputation," NBER Working Paper No. 2259, 1987.

______ and ______, "Sovereign Debt as a Contingent Claim: Excusable Default, Repudiation, and Reputation," *American Economic Review*, December 1988, *78*, 1088–97.

Holmström, Bengt and Myerson, Robert, "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica*, November 1983, *51*, 1799–819.

Ledyard, John "The Pure Theory of Large Two-Candidate Elections," *Public Choice*, August 1984, *44*, 7–41.

Lucas, Robert and Stokey, Nancy, "Optimal Monetary and Fiscal Policy in an Economy Without Capital," *Journal of Monetary Economics*, July 1983, *12*, 55–93.

Peleg, Bezalel and Yaari, Menahem, "On the Existence of a Consistent Course of Action When Tastes Are Changing," *Review of Economic Studies*, July 1973, *40*, 391–401.

Persson, Torsten and Svensson, Lars, "Checks and Balances on the Government Budget," *Quarterly Journal of Economics*, May 1989, *104*, 325–46.

Ratchford, B.V., *American State Debts*, Durham, NC: Duke University Press, 1941.

Rawls, John, *A Theory of Justice*, Cambridge: Harvard University Press, 1971.

Rogoff, Kenneth and Sibert, Anne, "Equilibrium Political Business Cycles," *Review of Economic Studies*, January 1988, *55*, 1–16.

Roubini, Nouriel and Sachs, Jeffrey, (1989a) "Political and Economic Determinants of Budget Deficits in Industrialized Democracies," *European Economic Review*, forthcoming, 1989.

______ and ______, (1989b) "Government Spending and Budget Deficits in the Industrial Countries," *Economic Policy*, Spring 1989, *8*.

Sachs, Jeffrey, "Theoretical Issues in International Borrowing," *Princeton Studies in International Economics*, 54, May 1984.

Scott, William A., *The Repudiation of State Debts*, Boston: T. Crowell & Co., 1893.

Strotz, Robert, "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, *23*, 1956, 165–80.

Tabellini, Guido, "Domestic Politics and the International Coordination of Fiscal Policies," *Journal of International Economics*, forthcoming, 1989.

______, "The Politics of Intergenerational Redistribution," NBER Working Paper No. 3058, 1989.

Arming as a Strategic Investment in a Cooperative Equilibrium

By MICHELLE R. GARFINKEL*

To develop a positive, economic theory of military spending, the analysis focuses on a game-theoretic, general equilibrium model of international conflict, in which consumption, peaceful investment, and military spending are endogenously determined. The analysis illustrates that when there is repeated interaction between nations, a game of threats and punishments generally will not support a disarmament outcome and that fluctuations in military spending can be an endogenous result of fluctuations in aggregate economic activity. Furthermore, the analysis shows how the relation between aggregate economic activity and military spending qualitatively depends on whether governments are acting opportunistically or cooperatively. (JEL 114,321)

This paper develops a positive theory of military spending on armaments. The analysis focuses on a game-theoretic, general equilibrium model of international conflict, in which resource allocation among consumption, peaceful investment, and military spending is endogenously determined. The analysis shows that fluctuations in military spending can be an endogenous result of fluctuations in aggregate economic activity. Accordingly, the analysis raises doubts about existing empirical studies of macroeconomic fluctuations in which military spending is treated as exogenous—for example, see Robert Hall (1986).[1] In addition, taking the repeated interaction between nations explicitly taken into account, the analysis shows that the qualitative relation between military spending and aggregate economic activity depends on whether governments are acting cooperatively or opportunistically. Specifically, if governments are acting cooperatively, then the model predicts that, for a given realization of aggregate economic activity, the associated level of military spending will be larger, the larger is the variance of aggregate economic activity and the smaller is the average level of economic activity. Alternatively, if governments were acting opportunistically, then the model predicts that for a given realization of aggregate economic activity, the associated level of military spending would be independent of the average level and the variance of aggregate economic activity.

In contrast to the existing literature on game-theoretic models of international conflict,[2] the game-theoretic analysis developed

---

*Economist, Federal Reserve Bank of St. Louis, St. Louis, MO 63166. I wish to thank Peter Garber, Herschel Grossman, Seonghwan Oh, an anonymous referee, and the participants of the NBER-FRG Summer Institute 1987 for helpful comments. The opinions expressed in this paper are those of the author and do not necessarily represent the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

[1] The existing empirical evidence on whether military spending is Granger-causal prior to other aggregate economic variables is mixed. For example, Teresa Garcia-Mila (1987) documents evidence that is consistent with the notion that military spending is statistically exogenous with respect to other aggregate economic variables. But, R. P. Smith (1987) finds that while past values of income and civilian expenditures can help to predict current military spending, past values of military spending do not help predict income and civilian expenditures.

[2] See Michael Intriligator (1982) for a list of references and Martin Shubik (1987) for a survey of the literature and a critical assessment of the applicability of game-theoretic techniques to the study of the issues in conflict theory. Early applications of the tools of game theory to the study of international conflict, Thomas Schelling (1960) and Intriligator (1975) for example, focus on the study of arms races. Subsequent research—see Dagobert Brito and Intriligator (1984), for example—extends the analysis to investigate the possibility of an arms race leading to a war. Researchers

in the present paper emphasizes two economic aspects of international conflict. First, there are the economic costs incurred by nations who choose to participate in international conflict. The economic costs arise most obviously through a nation's resource constraints. Although the existing analyses include a high degree of mathematical sophistication and address important issues, they are incomplete as they fail to account for resource constraints and, more generally, they abstract from the endogenous determination of resource allocation.

Second, there might be possible economic (among other) motives for nations to confront other nations. Lionel Robbins (1940), in assessing the extent to which war can be regarded as a result of economic motives, suggests as an example that a government acting to promote the interests of its citizens might go to war, whereby, if successful, it could improve its terms of trade and hence economic welfare.[3] Although tributes in this century generally have not been significant,[4] victories in confrontations historically have generated substantial tributes (especially for nations confronting weaker neighboring nations).[5]

Nevertheless, economists who have examined the game-theoretic aspects of economic policymaking in interdependent economies —see, for example, Matthew Canzoneri and Jo Anna Gray (1985) and Canzoneri and Dale Henderson (1986, 1987)—abstract from war, or a threat thereof, as a possible strategy to promote the economic interests of the nation. Dagobert Brito (1972), who develops a two-nation, general equilibrium model of an armaments race and accounts for the

---

in conflict theory also have created a large literature concerned with retaliation and deterrence. Building on the seminal work of Steven Brams and Donald Wittman (1981), the aim of recent research—for example, see Frank Zagare (1985) and Brams and Marc Kilgour (1985, 1986)—has been to identify and to characterize the nonmyopic equilibria in abstract settings in which the players, interpreted as nations, can make sequential moves conditional on the initial move of their opponent. The equilibrium outcome is finally determined when none of the players has the desire to make a further move. Robert Axelrod (1984) finds that "tit-for-tat" strategies, in which retaliation is possible because of the repeated nature of the game, can explain why there is not a perpetual worldwide war.

[3] Recently, Tom Riddell (1988), who recognizes the importance of the link between a nation's international economic objectives and its military spending, documents a positive relation between U.S. military power and its economic welfare in the post-World War II period.

[4] Even in cases where nations do not appear to be better off as a result of wartime activities, it is not clear that these nations did not participate in hope of obtaining tribute. However, see William Shepherd (1988), who provides an interesting discussion of the relevance of the tribute as a motivation for military spending as a defensive as well as an offensive measure. As he notes, there have been instances of large tributes in this century. For example, Alan Milward (1970, p. 273) calculates that payments to Germany from France as a percentage of French national income in 1938 at 1938 prices from 1940 through the first six months of 1944 were respectively 10.9, 19.3, 20.9, 36.6, and 27.6—an average of 23 percent of 1938 French national income per year. But, taking into account the costs of the German war machine (even excluding casualties) and those of occupation, tends to diminish the relevance of the tribute.

[5] Whether or not the ability to extract or to avoid paying tribute provides a sufficient motivation to endure the costs of producing armaments is a matter of controversy. Indeed, it seems reasonable to argue that most modern conflicts between nations are driven not by booty, but by differences in religion or ideologies, for example. Although such differences are important in international conflicts, however, it is not clear that they serve as the primary driving force for confronting other nations. For example, following Douglas North (1981), who argues that ideologies are crucial for circumventing free-rider problems, one could argue that ideological differences might emerge as a government's justification to build its military strength. Without ideologies, individual citizens would not willingly contribute to the nation's mobilization efforts, since it is not feasible for the government to make the individual's return on the military investment as a function of his own contribution. Hence, while differences in ideologies might appear to be at the root of most modern international conflicts, such differences actually might emerge as a by-product of a nation's mobilization efforts to extract tribute from other nations or avoid paying tribute to stronger nations. Alternatively, tribute could be thought of as the gain in utility generated by the success of a nation's wartime efforts, though the actual transfer of goods might be negligible or even negative. The analysis to follow could incorporate such a notion of tribute. Provided that consumption goods and the success in an ideological or religious dispute are relatively close substitutes, the qualitative results to follow would hardly change.

costs of such a race through a nation's resource constraints, does not identify the benefits of such a race. Instead, he simply includes defense expenditures by both nations in individuals' utility function, presumably to capture the notion that such expenditures, holding the enemy's expenditures fixed, generate psychic benefits.[6] Joshua Feinman, Peter Garber, and Michelle Garfinkel (1986) develop a general equilibrium model in which nations arm in order to advance the economic interests of their citizens, who do not directly derive utility from armaments. The result in this model that economic motives lead to an arms race is not general, however, as this model assumes that the governments play a one-shot game.

The model in this paper indirectly builds on an economic motivation for investing in armaments, but formalizes the notion that modern conflicts, as reflected in a nation's military spending, might not be directly about tribute. Success in a confrontation, which depends on the relative amounts of armaments each nation constructs, would generate goods for future consumption. As in Feinman, Garber, and Garfinkel (1986), without repeated interaction between the players, in which case governments act noncooperatively by taking the opponent's strategy as given, a significant amount of resources would be wasted away in military spending. Although no transfer of goods would be observed in this hypothetical outcome, each nation's choice to endure the costs of arms production would be driven by the possibility of extracting or avoiding extraction of resources. In the one-shot game, a commitment to peaceful policies, with perhaps a voluntary transfer of goods, is not an equilibrium strategy. Specifically, because each government would choose to deviate from the commitment policy by producing armaments, disarmament is not a subgame perfect equilibrium.

With repeated interaction between sovereign states, however, threats and punishments can support cooperative behavior.[7] The resulting "cooperative equilibrium," where cooperation is not assumed but derived as an optimal strategy, corresponds to what is referred to as the "trigger strategy equilibrium" in the game theory literature. In the cooperative equilibrium, disarmament is a possibility. More generally, the amount of resources allocated to military spending is positive, possibly with intermittent periods of positive and zero military spending on armaments. But, for each play of the game, military spending is less than in the one-shot game.

Military spending in the cooperative equilibrium does not arise directly from an economic motive, as in the one-shot game without commitments. Instead, the positive production of armaments arises from a deterrence motive, for it is necessary to induce cooperative behavior.

In what follows, the next section presents a modest extension of the economic model of international conflict developed in Feinman et al. (1986) and examines the characteristics of the equilibria that would emerge if the governments were to act "opportunistically" (the opportunistic equilibrium) and, alternatively, if governments were able to commit themselves to disarmament (the efficient outcome). Section II examines the role of threats and punishments in the cooperative equilibrium, identifying the conditions under which the strategic interplay between governments would be sufficient to support the (otherwise time-inconsistent) disarmament solution. Furthermore, this section characterizes military spending in the cooperative equilibrium when threats and punishments are not sufficient to support disarmament, and derives the qualitative differences between military spending in the opportunistic and the cooperative equilibria. The qualitative differences, in turn, imply some testable hypotheses to determine whether governments are acting opportunistically or cooperatively. Finally, Section III summarizes the results.

---

[6] See also F. van der Ploeg and A. J. de Zeeuw (1986a, b). Such a formulation might be made consistent with the idea mentioned above that tribute could be the gain in utility from a successful confrontation driven by ideological or religious differences.

[7] Mancur Olson and Richard Zeckhauser (1966), who analyze a model of alliances to formulate a theory of military spending that can explain the fact that larger members typically devote a larger proportion of available resources to military spending than smaller members, assume that members of the alliance act opportunistically. They implicitly assume, however, that there is sufficient cooperation for the existence and stability of the international organization. Moreover, their analysis ignores the optimizing behavior of nonmember nations, and in particular, the influence that enemy nations' military spending has on the effectiveness of member nations' military spending in providing defense for the alliance.

I. Analytical Framework

This section presents a simple two-sovereign state model of international conflict. As in most analyses of international conflict, this analysis takes the political structure as given. That is, it is assumed that the problems of coordination cannot be resolved by combining the two states into one. The model is essentially an infinitely repeated game version of the two-period war game originally discussed in Feinman et al. (1986). In particular, the simple two-sovereign state, two-period, general equilibrium model discussed in that paper is placed in a framework in which infinitely lived governments maximize the utility of their citizens who are also infinitely lived. Individuals in the domestic state and the foreign state, denoted by "*," have identical tastes, technologies, and endowments. Every two periods, $t = 0, 2, 4, \dots, \infty$, each individual receives an endowment of Z_t units of non-storable goods, which he allocates to current consumption, c_t, and to the production of goods to be consumed in the next period, c_{t+1}. Although Z_t is realized before individuals make their decisions in period t, future values of the endowment are uncertain. The endowment is generated by a stationary stochastic process, whose distribution is given by $F(Z)$ and $\int_{Z_l}^{Z_h} dF(Z) = 1$. Let $\bar{Z}$ denote the expected value of the endowment and σ^2 denote the variance of the endowment.

Individuals maximize expected lifetime utility, which depends on consumption in all periods of their lives. The unconditional expectation of a representative individual's lifetime utility is given by

$$(1) \quad E\{U_0\} = \sum_{t=0}^{\infty} \int_{Z_l}^{Z_h} \beta^t u(c_t)\, dF(Z), \qquad 0 < \beta < 1,$$

where β reflects the subjective rate of time preference and $E\{\cdot\}$ denotes the unconditional expectations operator. $u(c_t)$ satisfies $u'(c_t) > 0$, $u'(0) = \infty$, and $u''(c_t) < 0$.

There are two ways in which the endowment good can be transformed into goods for future consumption.[8] The first method is a peaceful investment technology, $f(i_t)$, with $f(0) = 0$, $f'(i_t) > 0$ and $f''(i_t) < 0$, enabling an individual to transform i_t units of his endowment received in t into $f(i_t)$ goods for consumption in the next period, $t+1$, for $t = 0, 2, 4, \dots, \infty$.

The second method of transferring endowment goods received in t to $t+1$ involves military spending. A unit of the endowment good can be converted into a unit of war goods, W_t, which is also non-storable.[9] The possibility of extracting resources from an enemy state can serve as a motivation for producing war goods (or equivalently armaments).[10] In this case, although war goods do not directly yield utility to individuals, they can do so indirectly, as success or failure in the confrontation results in an increase or a decrease in goods for consumption in the next period, through the imposition of a tribute on the weaker nation.

---

[8] For simplicity, it is assumed that these methods, which generate a certain return, are only available in the periods in which the endowment is received. While the qualitative results do not hinge on this simplification, this assumption, in conjunction with the assumptions that $Z_{t+1} = 0$ and that goods are non-storable (so that there is no accumulation of capital stocks), is important for it strips the model of any dynamics that would otherwise obscure the dynamics introduced by strategic considerations. Also, the analysis abstracts from labor as a factor of production in both methods. Including labor would not change the analysis to follow substantially.

[9] Ruling out the possibility of building up a capital stock of armaments with military spending makes the model tractable, and, as noted above, permits the analysis to highlight the dynamics arising from strategic considerations.

[10] However, as shown below in Section II, in the cooperative equilibrium, such a possibility serves only as an indirect motive for producing war goods. Specifically, the primary motivation is to induce cooperation by reducing each nation's incentive to act opportunistically in an effort to extract resources from the enemy nation.

The outcome of a confrontation, which is reflected in the amount of tribute one nation extracts from the other nation, depends on the relative amounts of war goods constructed by the two nations. Specifically, tribute per citizen of the domestic nation in $t+1$, which can be positive or negative, is determined by:

$$(2)\quad \Gamma(W_t, W_t^*) = \begin{cases} a(W_t - W_t^*), & \text{if } -f(i_t) \le a(W_t - W_t^*) \le f(i_t^*); \\ f(i_t^*), & \text{if } a(W_t - W_t^*) > f(i_t^*); \\ -f(i_t), & \text{if } a(W_t - W_t^*) < -f(i_t), \end{cases}$$

for $t = 0,2,4,\dots,\infty$. The expression in (2) says that the amount of tribute extracted depends linearly on the difference in the amount of war goods constructed by the two nations, but it cannot exceed the amount of goods available to the weaker state in $t+1$. The marginal product of war goods, a, is strictly positive. As noted below, a stable equilibrium requires $a < f'(0)$.

Different magnitudes of a gives rise to different possible interpretations of the tribute function.[11] If, for instance, a is very large, then the expression in (2) would capture the case in which one nation, having one more armament than the opponent, could extract the opponent's entire $t+1$ return on the peaceful investment.[12] Alternatively, if a is not very large, then the tribute function can be thought of as the result of a bargaining process between the two nations. War goods, then, enhance each nation's bargaining power according to (2). For example, in a more elaborate model, the weaker nation would be forced to coordinate its macroeconomic policy in a way that improves the economic welfare of the stronger nation and simultaneously deteriorates the economic welfare of the weaker nation.

In any case, (2) assumes that the outcome of the confrontation simply involves a comparison of the amounts of war goods constructed by each nation. Note that the expression in (2) implies that the returns from military spending for the domestic and the foreign states must sum to zero—that is, $\Gamma(W_t, W_t^*) = -\Gamma^*(W_t^*, W_t)$.[13]

Assume that each of the infinitely lived governments of the two nations gains the resources necessary to play the war game through lump sum taxation of its citizens in t and distributes or raises the tribute through lump sum transfers or taxes in $t+1$ for $t = 0,2,4,\dots,\infty$. Assume also that governments' policies are formulated in the interests of the representative individual.[14] Accordingly, at the beginning of every two periods, each government selects the magnitude of military spending, W_0, and plans a sequence of future contingent levels of military spending, $\{W_t\}_{t=2,4}^{\infty}$ to maximize the expected utility of the representative individual. The domestic government maximizes the objective function given by (1) subject to (2) and the following constraints:

---

[11] See Shepherd (1988) who describes the possible technological and institutional factors that determine the magnitude of the parameter a. Another interpretation of the conflict as embedded in (2), would arise in a nonnation player setting. For example, the players might be pirates or thieves. With minor modifications, the model also might apply to the game of contestable transfers, where groups lobby in the political arena to change the existing policy. A group's failure to participate could be costly if that group would have to help to finance that transfer to the winner.

[12] The value of a that would permit such an interpretation can be finite.

[13] One possible extension of the model that would permit a clear distinction between mobilization and actual warfare would involve assuming that the endowment can differ *ex post* across nations and that both technologies for transferring goods into the future are stochastic. In particular, suppose that the return from the peaceful investment were realized in $t+1$, for $t = 0,2,\dots,\infty$, and could differ *ex post* across nations. At that time, after mobilization decisions have been made, each nation decides whether or not to initiate a war, the return from which is uncertain. The variable that would capture the uncertainty associated with fighting, in the form of an additive term in the $t+1$ resource constraint (see equation (3b)), would be identically zero if there were no fighting and strictly negative for each nation initiating a strike. In this extension, involving a two-stage maximization problem, warfare would not be a zero sum game.

[14] Principal-agent problems are ignored.

$$c_t = Z_t - i_t - W_t \tag{3a}$$

$$c_{t+1} = f(i_t) + \Gamma(W_t, W_t^*), \tag{3b}$$

for $t = 0, 2, 4, \ldots, \infty$. The foreign government solves an analogous problem.

Parts A and B study the opportunistic and the efficient solutions, which differ with respect to the specification of a government's strategy and perception of its opponent's strategy. These hypothetical solutions serve as benchmark solutions to which the cooperative equilibrium, studied below in Section II, can be compared. In the analyses to follow, it is assumed that the informational structure of the model is perfect in the sense that each government knows the opponent's optimization problem and the constraints (including the current realization of the other state's endowment) faced by the opponent. In addition, it is assumed that the governments act simultaneously. This assumption is sufficient to ensure perfection in the one-shot game.

A. *The Opportunistic Equilibrium*

This section assumes that each government solves its problem of maximizing the utility of the representative agent (1), taking its opponent's military spending as given. The resulting equilibrium—the opportunistic equilibrium—corresponds to the subgame perfect Nash (noncooperative) equilibrium in the infinite horizon game, assuming that a government's strategy is such that its actions are in no way contingent on the previous actions of the opponent. Such an equilibrium is unsatisfactory, for it effectively assumes that the governments do not recognize the impact that their current policy can have on their opponents' future policies and hence on future outcomes. Such an assumption, which would be realistic only if individuals and governments live for two periods, leads to the equilibrium studied by Feinman et al. (1986).

At the beginning of every two periods, $t = 0, 2, \ldots, \infty$, each government chooses levels of peaceful investment and military expenditures to maximize the utility of the representative agent (1), subject to the tribute function (2) and the resource constraints (3), given the current realization of the endowment and taking the opponent's strategy as given. The assumptions imposed on the instantaneous utility function, $u(c_t)$, and the peaceful investment technology, $f(i_t)$, imply that the solution to the governments' problem must satisfy the following first-order condition:

$$\frac{u'(c_t)}{u'(c_{t+1})} = \beta f'(i_t), \tag{4a}$$

for $t = 0, 2, \ldots, \infty$, where $f'(0) > a$.[15] In addition, assuming an interior solution for military spending in period t, which implies that $f'(Z_t) < a$, the solution must satisfy the following first-order condition:

$$f'(i_t) \leq a, \tag{4b}$$

for $t = 0, 2, \ldots, \infty$. The expression in (4b) possibly results in a strict inequality condition because of the truncation of the tribute function.

In order to derive specific functional forms for the reaction functions, which relate a nation's military spending to that of the opponent, the following specifications for utility and the peaceful investment technology are made:

---

[15] If $f'(0) \leq a$, then peaceful investment in period $t = 0$ would be zero for both nations, so that investment in armaments would be fruitless. But, given that the opponent produces zero armaments, each government would be better off by investing some fraction of the endowment in the peaceful technology, thereby giving rise to the opponent's incentive to shift resources from the peaceful technology to the production of armaments. The oscillating nature of this reaction process suggests that a stable Nash equilibrium does not exist in the one-shot game when $f'(0) \leq a$.

$$(5\text{a}) \qquad v_t = \ln(c_t) + \beta\ln(c_{t+1})$$

$$(5\text{b}) \qquad f(i_t) = Ai_t^\alpha, 0 < \alpha < 1,$$

for $t = 0,2,4,\dots,\infty$, where v_t denotes the representative agent's two-period utility starting in period t.[16] Given this specification, the first-order conditions from choosing optimal investment and military expenditures, given by (4), imply the following reaction function for the domestic government:

$$(6) \quad \overline{W}_t(W_t^*) = \frac{\beta}{1+\beta}Z_t - \frac{1+\alpha\beta}{\alpha(1+\beta)}\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}} + \frac{1}{1+\beta}W_t^*,$$

for $t = 0,2,4,\dots,\infty$. The reaction function for the foreign government is found by simply replacing W_t^* with W_t. As is evident from (6), a nation's military spending is positively related to that of its opponent, but this relationship is not one for one. Generally speaking, provided that consumption goods, c_t and c_{t+1}, are normal, an increase in an opponent's military spending will induce a nation to increase its own military spending, but by less than the increase in the opponent's spending.

If $Z_t > ((1+\alpha\beta)/\alpha\beta)(\alpha A/a)^{1/(1-\alpha)} \equiv k$, as assumed throughout the analysis below, each period $t = 0,2,\dots,\infty$, both governments will construct positive amounts of war goods. By combining the domestic government's reaction function (6) with that for the foreign government, noting that the assumption of identical nations implies $\overline{W}_t = \overline{W}_t^*$, and using (3a), (3b), and (5), one can establish that the solutions to this game are given by

$$(7\text{a}) \qquad \overline{W}_t = Z_t - k$$

$$(7\text{b}) \qquad i_t = \frac{\alpha\beta}{1+\alpha\beta}k$$

$$(7\text{c}) \qquad c_t = \frac{1}{1+\alpha\beta}k$$

$$(7\text{d}) \qquad c_{t+1} = \frac{a\beta}{1+\alpha\beta}k,$$

for $t = 0,2,4,\dots,\infty$. In addition to there being a stalemated confrontation, symmetry implies that peaceful investment and consumption are identical in the foreign state. In this equilibrium, the utility of the domestic government is given by the following:

$$(8) \qquad U_0^n = \frac{1}{1-\beta^2}\left((1+\beta)\ln\frac{k}{1+\alpha\beta} + \beta\ln a\beta\right)$$

and symmetry implies that $U_0^n = U_0^{n*}$.[17]

Because war goods do not directly yield utility to individuals, the amount of resources allocated to military spending represents a deadweight loss arising from the international conflict. In the opportunistic equilibrium, $\overline{W}_t$ goods are essentially thrown away in periods $t = 0, 2, \ldots, \infty$ and any increase in the endowment above k is simply wasted away in military spending.

---

[16] Note that this specification for the peaceful investment technology ensures that $f'(0) > a$ for any finite value of a, such that the analysis to follow is robust to the different interpretations of the tribute function suggested above. Furthermore, note that the specifications made in (5) can be thought of as an approximation to a more general set of functions.

[17] As indicated by (7a), provided that the endowment is sufficiently large, any increase in the endowment above k is allocated to military spending. This result depends on the constancy of the marginal product of war goods. Even if the marginal product of war goods were diminishing, some positive fraction of the hypothetical increase in the endowment would be allocated to the war effort. The marginal product of war goods is assumed to be constant to highlight the tradeoff between peaceful and military investment, and thereby the impact of the strategic behavior of governments to be studied below in Section II.

B. *The Efficient Outcome*

In the section above, the equilibrium concept—that is, the one-shot, Nash equilibrium—effectively assumes that governments ignore the potential impact that they can have on the enemy government's policies. In contrast, this section imagines that governments are able to coordinate their policies and to effectively commit themselves to current and future policies regarding military spending. In such a scenario, disarmament prevails, and the waste of resources is eliminated: $W_t = W_t^* = 0$ for all t.

The efficient outcome is the solution to the maximization of (1), subject to (3), given the current realization of the endowment, Z_0, and $W_t = W_t^*$ for all t. The first-order condition (4b) becomes irrelevant. The solutions in the efficient outcome are given by the following:

$$\text{(9a)} \qquad i_t = \frac{\alpha\beta}{1+\alpha\beta} Z_t$$

$$\text{(9b)} \qquad c_t = \frac{1}{1+\alpha\beta} Z_t$$

$$\text{(9c)} \qquad c_{t+1} = A\left(\frac{\alpha\beta}{1+\alpha\beta} Z_t\right)^{\alpha},$$

for $t = 0, 2, 4, \ldots, \infty$. As indicated by (9), $W_t = W_t^* = 0$. Symmetry implies again that consumption and investment in the foreign state are also given by (9). Consequently, the unconditional expected utility for the domestic government, which is identical to that for the foreign government, is given by the following:

$$\text{(10)} \quad E\{U_0^c\} = \sum_{t=0,2}^{\infty} \int_{Z_l}^{Z_h} \beta^t \left((1+\alpha\beta)\ln\frac{Z_t}{1+\alpha\beta} + \beta \ln A + \alpha\beta \ln \alpha\beta \right) dF(Z).$$

By subtracting (8) from (10), one can evaluate the expected difference between the utility obtained in the efficient solution and that obtained in the opportunistic equilibrium:

$$\text{(11)} \quad E\{U_0^c - U_0^n\} = \sum_{t=0,2}^{\infty} \int_{Z_l}^{Z_h} \beta^t (1+\alpha\beta) \ln \frac{Z_t}{k} > 0.$$

The expected difference is strictly positive if the endowment, Z_t, is sufficiently large to ensure positive military expenditures in the opportunistic equilibrium for periods $t = 0, 2, \ldots, \infty$.

Without repeated interaction between governments, a third party or some institution to enforce commitments, neither state could avoid the temptation to abrogate the commitment to zero military spending. The domestic government, for example, would have an incentive not to cooperate by constructing a positive amount of war goods according to the reaction function (6), $\overline{W}(W^* = 0, Z_t)$. Of course, the foreign government would have the same incentive. In other words, in the one-shot game, the absence of enforceable commitments to disarmament precludes the feasibility of the disarmament outcome. However, as demonstrated in the next section, with repeated interaction, threats and punishments might be able to support the disarmament outcome.

II. Threats, Punishments, and Cooperation

This section contains an examination of strategic considerations which can substitute, at least to some degree, for each government's inability to make commitments. The fundamental idea here is that when nations repeatedly interact, each government can indirectly influence its opponent's strategy with threats and punishments for deviant behavior, without having to rely on commitments. Threats and punishments create a link between a government's past policy and its opponent's current policy. This link provides the mechanism whereby a government can influence its opponent's future policies through its own current actions. If, for example, the foreign government cooperates, by pursuing the endogenously determined cooperative policy, then the domestic government will implement the cooperative policy in the next play of the game. If, however, the foreign government acts opportunistically, then the domestic government imposes some punishment, the threat of which must be credible. Because the threat of punishment, triggered by opportunistic behavior, serves as sufficient motivation for both governments to maintain cooperation, the punishment is never observed in the equilibrium.

The cooperative equilibrium considered in this section assumes that the governments employ the trigger strategy originally studied by James Friedman (1971). Trigger strategies, which differ with respect to the length of the punishment, specify that the punishment for deviating from the endogenously determined cooperative policy involves a reversion to the opportunistic solution. While such trigger strategies, as an expectations mechanism, might be realistic in models of monetary policy, such as Robert Barro and David Gordon (1983) and Herschel Grossman and John Van Huyck (1986), where the government is playing a game against atomistic agents, limiting consideration to trigger strategies in the context of this model be questionable. Specifically, as demonstrated by Dilip Abreu (1986), if no trigger strategy is sufficiently severe to support the peaceful outcome, then more severe punishment strategies could be considered. Presumably, governments acting optimally announce and follow strategies that involve the most severe, credible threat, because the corresponding punishment strategy will lead to the best outcome possible. As argued at the end of this section, however, the result that threats and punishments cannot always support the disarmament solution does not depend on the assumption that the governments announce and follow a trigger strategy.

Assume that in period $t=0$, the domestic government expects the foreign government to follow a contingent cooperative policy, denoted by $\tilde{W}(Z_t)$, for $t=0$ and so announces that it too will follow this contingent cooperative policy in $t=0$. Furthermore, the domestic government announces that it will continue to follow this contingent cooperative policy for $t=2,4,\dots,\infty$, provided that the opponent has never deviated from that policy. The domestic government threatens to act opportunistically forever if the foreign government ever acts opportunistically. Specifically, the domestic government announces the following strategy:

$$W_0=\tilde{W}(Z_0),$$

$$(12)\quad W_t=\begin{cases}\tilde{W}(Z_t), & \\ \quad \text{if } W_s^*=\tilde{W}^*(Z_s) & \\ \qquad \text{and} \quad W_s=\tilde{W}(Z_s), & \\ \qquad 0\leq s<t; & \\ \overline{W}_t(\overline{W}_t^*,Z_t), & \\ \quad \text{otherwise,} & \end{cases}$$

for $t=2,4,\dots,\infty$.[18] The foreign government announces a similar strategy. According to this strategy, the punishment period for opportunistic behavior is infinite. While a shorter punishment could be negotiated by the two governments, the strategy given by (12) is subgame perfect, for it is simply the best one-shot, optimal response.[19] Moreover, among the whole set of trigger strategy equilibria, the solution that involves an infinite punishment period supports the best outcome. Thus, the trigger strategy described in (12) is optimal within the class of trigger strategies.

---

[18] In the extension of the model, briefly described in fn. 13, to permit a distinction between mobilization and warfare, the deterrence effect of military spending embedded in the trigger strategy could be twofold. First, as in the present model, it would reduce the incentive to build a war machine in an ultimate effort to extract tribute. Second, given each nation's mobilization efforts and the realization of the return on the peaceful investment (both of which can differ across nations), military spending by each nation would reduce the incentive of the enemy to engage in a preemptive strike by providing the armaments for a retaliatory strike by the potentially preempted nation.

[19] One possible objection that can be made against the trigger strategy, as well as the strategies studied by Abreu (1986), revolves around the possibility of future renegotiation. Specifically, the subgame perfection property does not preclude the possibility that once one government has deviated from the cooperative policy, both governments might be able to agree to skip the punishment phase of the strategy and start the game over. If both governments recognize this possibility, then a new situation arises and the credibility of the punishment is questionable. One potential way to treat this problem is to model the probability of successful

The problem for the two governments in period $t=0$ is to choose $\tilde{W}(Z_0)$ and the contingent cooperative policy, $\tilde{W}(Z_t)$ for $t=2,4,\dots,\infty$, so as to maximize the expected utility of their representative citizens (1), subject to (3) and (12). The choice of the contingent cooperative policy defines the cooperative equilibrium, in which for all possible realizations of the endowment, Z_0, neither government has an incentive to act opportunistically in period $t=0$ nor expects to have such an incentive in periods $t=2,4,\dots,\infty$. The optimal cooperative policy, $\tilde{W}(Z_t)$, is a member of the set of possible cooperative solutions. As a member of this set, $\tilde{W}(Z_t)$ must satisfy the following subgame perfection constraint for all possible values of Z_0:

$$(13)\quad v_0\big(\overline{W}(\tilde{W}^*(Z_0)),\tilde{W}^*(Z_0)\big) - v_0\big(\tilde{W}(Z_0),\tilde{W}^*(Z_0)\big) \le \beta^2 E\{\tilde{U}_2 - U_2^n\}$$

where

$$v_i(W,W^*,Z) = v_i\big(c_i(W,Z),\; c_{i+1}(W-W^*,Z)\big),$$

$$E\{\tilde{U}_2\} = \sum_{i=2,4}^{\infty}\int_{Z_l}^{Z_h}\beta^{i-2} \times v_i\big(\tilde{W}(Z_i),\tilde{W}^*(Z_i),Z_i\big)\,dF(Z_i),$$

$$E\{U_2^n\} = \sum_{i=2,4}^{\infty}\int_{Z_l}^{Z_h}\beta^{i-2} \times v_i\big(\overline{W}(Z_i),\overline{W}^*(Z_i),Z_i\big)\,dF(Z_i),$$

and $\overline{W}(\tilde{W}^*(Z_0), Z_0)$ denotes the opportunistic amount of military spending by the domestic government, given that the foreign government is cooperating in period $t=0$. A similar condition must hold for the foreign nation. Given the stationarity of the stochastic process generating Z_t, the solution, $\tilde{W}(Z_t)=\tilde{W}^*(Z_t)$, that satisfies (13), will satisfy a condition analogous to (13) for $t=2,4,\dots,\infty$.

The left-hand side of (13) is equal to the "temptation" to not cooperate. It is the difference in the two-period utility obtained when acting opportunistically and that obtained when implementing the cooperative policy. The temptation will be strictly positive for $\tilde{W}(Z_0) < \overline{W}(\tilde{W}^*, Z_0)$. The right-hand side of the inequality (13) is the expected difference between utility obtained when implementing the cooperative policy and that obtained when implementing the opportunistic policy starting in $t=2$. In other words, it is government's expected gain in utility from continuing to cooperate and foregoing the immediate gain that could be realized by failing to cooperate in period $t=0$. In contrast to the temptation, which depends on the current realization of the endowment, the expected gain is fixed. The subgame perfection constraint (13) requires that the fixed expected gain exceed the temptation for all possible realizations of the endowment.

There is at least one solution ($\tilde{W}(Z_t)$, $\tilde{W}^*(Z_t)$) that satisfies (13)—namely, the opportunistic solution, which makes both sides of the inequality in (13) equal to zero. Depending on the parameters of the model, including the distribution of the endowment, the peaceful investment technology, the tribute function, and the discount factor, the set of possible cooperative equilibria, which will generally have more than one element, might include the disarmament solution. In any case, given that the governments are following the trigger strategy specified in (12), because the amount of resources allocated to military spending in cooperative equilibrium represents a deadweight loss, the optimal contingent cooperative policy, $\tilde{W}(Z_t)$, will specify the smallest, incentive compatible amount of military spending for each real-

---

renegotiation, in the case that one government has behaved opportunistically, as an exogenous stochastic process. (See Grossman and Van Huyck (1988) who similarly model the memory of private agents as a stochastic process) ization of the endowment—that is, it will equal the smallest $\tilde{W}(Z_t)$ that satisfies (13).

A. *Incentive Compatible Military Spending*

In the example in which preferences and the peaceful investment technology are given by (5), the trigger strategy generally does not serve as a perfect substitute for the ability to make commitments. To see that the disarmament solution is not necessarily sustained in a trigger strategy equilibrium, it is helpful to calculate the temptation to deviate from the cooperative policy and the gain from maintained cooperation for the cooperative policy $\tilde{W}(Z_t)$.

Given that the foreign government cooperates in $t=0$, the opportunistic behavior by the domestic government, using the reaction function (6), the first-order conditions (4) and the budget constraints (3), are given by the following:

$$\text{(14a)}\quad \overline{W}(\tilde{W}(Z_0),Z_0)=\frac{\beta}{1+\beta}(Z_0-k)+\frac{1}{1+\beta}\tilde{W}(Z_0)$$

$$\text{(14b)}\quad i_0=\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}}$$

$$\text{(14c)}\quad c_0=\frac{1}{1+\beta}\times\left(Z_0+\frac{1-\alpha}{\alpha}\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}}-\tilde{W}(Z_0)\right)$$

$$\text{(14d)}\quad c_1=\frac{\alpha\beta}{1+\beta}\times\left(Z_0+\frac{1-\alpha}{\alpha}\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}}-\tilde{W}(Z_0)\right),$$

where, as defined previously,

$$k\equiv\frac{1+\alpha\beta}{\alpha\beta}\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}},$$

the threshold level of Z above which there is military spending in the opportunistic equilibrium. The solution (14) assumes that the magnitude of the opportunistic level of military spending does not trigger the kink in the tribute function (2).[20] Such opportunistic behavior yields the following two-period utility:

$$\text{(15)}\quad v_0(\overline{W}(\tilde{W}(Z_0)),Z_0-\tilde{W}(Z_0))=(1+\beta)\ln\left(Z_0+\frac{1-\alpha}{\alpha}\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}}-\tilde{W}_0(Z)\right)\frac{1}{1+\beta}+\beta\ln\beta a.$$

Under the cooperative policy, both governments construct $\tilde{W}(Z_0)$. They maximize the representative agent's utility (1) with respect to peaceful investment, subject to (3) and $W_t=\tilde{W}Z_t$ for $t=0,2,\dots,\infty$. By using (4a) and the constraints in (3), one can easily verify that the optimal levels of consumption and the peaceful investment, for the cooperative policy, $\tilde{W}$, are given by

$$\text{(16a)}\quad i_t=\frac{\alpha\beta}{1+\alpha\beta}(Z_t-\tilde{W}(Z_t))$$

$$\text{(16b)}\quad c_t=\frac{1}{1+\alpha\beta}(Z_t-\tilde{W}(Z_t))$$

$$\text{(16c)}\quad c_{t+1}=A\left(\frac{\alpha\beta}{1+\alpha\beta}(Z_t-\tilde{W}(Z_t))\right)^{\alpha},$$

---

[20] That is to say, the following condition must hold:

$$\frac{\alpha\beta}{1+\beta}(Z_t-k-\tilde{W}(Z_t))\le A\left(\frac{\alpha\beta}{1+\alpha\beta}(Z_t-\tilde{W}(Z_t))\right)^{\alpha},$$

for $t=0$. In the case that this condition is not satisfied, the opportunistic behavior for the construction of war goods is on the kink of the tribute function,

$$\overline{W}(\tilde{W}(Z_t),Z_t)=\frac{A}{a}\left(\frac{\alpha\beta}{1+\alpha\beta}(Z_t-\tilde{W}(Z_t))\right)^{\alpha}.$$

The values of the optimal consumption and investment become complicated without providing any additional insight into the nature of the cooperative solution.

for $t = 0, 2, 4, \ldots, \infty$. Such cooperative behavior yields the following two-period utility:

$$(17)\quad v_0\big(\tilde{W}(Z_0), \tilde{W}(Z_0)\big) = (1+\alpha\beta)\ln\frac{Z_0 - \tilde{W}(Z_0)}{1+\alpha\beta} + \beta\ln A + \alpha\beta\ln\alpha\beta.$$

The difference between (15) and (17) represents the domestic government's temptation to deviate from the cooperative policy, contingent on the current realization of the endowment, given that the foreign government is cooperating. The temptation, which is an increasing function of the difference between the level of the endowment and the amount of cooperative military spending, equals zero for $\tilde{W}(Z_0) = \overline{W}(\overline{W}^*, Z_0)$ and is strictly positive for $\tilde{W}(Z_0) < \overline{W}(\overline{W}^*, Z_0)$. Letting $T(Z_0 - \tilde{W}(Z_0))$ denote the temptation to deviate from the cooperative policy, it is easy to establish that

$$\frac{\partial T\big(Z_0 - \tilde{W}(Z_0)\big)}{\partial \tilde{W}(Z_0)} = -\frac{(1+\beta)\dfrac{\alpha}{1-\alpha}}{\dfrac{\alpha}{1-\alpha}\big(Z_0 - \tilde{W}(Z_0)\big) + \left(\dfrac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}}} + \frac{1+\alpha\beta}{Z_0 - \tilde{W}(Z_0)}.$$

For $\tilde{W}(Z_0) < \overline{W}(\overline{W}^*, Z_0))$, an increase in the amount of cooperative war expenditures decreases the temptation to cheat, thereby serving as a deterrent against opportunistic behavior. Furthermore, it is easy to verify by differentiating the temptation with respect to Z_0, while holding $\tilde{W}_0$ fixed, that

$$\frac{\partial T\big(Z_0 - \tilde{W}(Z_0)\big)}{\partial Z_0} = -\frac{\partial T\big(Z_0 - \tilde{W}(Z_0)\big)}{\partial \tilde{W}(Z_0)}.^{21}$$

Thus, for cooperative levels of military spending below the opportunistic equilibrium level, the temptation is increasing in Z_0, the current realization of the endowment.

In contrast to the temptation, the expected gain is independent of the current realization of the endowment. The expected gain from maintained cooperation, denoted by $G(Z - \tilde{W}(Z))$, is found by using (10) and (17):

$$(18)\quad G\big(Z - \tilde{W}(Z), F(Z)\big) = \sum_{t=2,4}^{\infty}\int_{Z_1}^{Z_h}\beta^t(1+\alpha\beta) \times \ln\frac{Z - \tilde{W}(Z)}{k}\,dF(Z).$$

Again, for $\tilde{W}(Z_t) = \overline{W}(\overline{W}^*, Z_t)$, the gain equals zero, and for $\tilde{W}(Z_t) < \overline{W}(\overline{W}^*, Z_t)$, the gain is strictly positive. Furthermore,

$$\frac{\partial G\big(Z - \tilde{W}(Z), F(Z)\big)}{\partial \tilde{W}(Z)} = -\sum_{t=2,4}^{\infty}\int_{Z_1}^{Z_h}\beta^t(1+\alpha\beta) \times \frac{1}{Z - \tilde{W}(Z)}\,dF(Z) < 0.$$

---

[21] This result does not depend on the specifications made for the tribute function, (2), or for utility and the peaceful investment technology, (5). Specifically, using (1), (3), and (13) with the envelope theorem, one can easily verify the following:

$$-\frac{\partial T(\tilde{W}, Z)/\partial Z}{\partial T(\tilde{W}, Z)/\partial \tilde{W}} = -\frac{u'[Z - i - \overline{W}] - u'[Z - i - \tilde{W}]}{-\beta\Gamma'(\overline{W}, \tilde{W}^*)u'[f(i) + \Gamma(\overline{W}, \tilde{W}^*)] + u'[Z - i - \tilde{W}]}.$$

By virtue of the first-order condition, from the maximization of (1) subject to (3) with respect to W, the right-hand side of the above expression simplifies to 1.

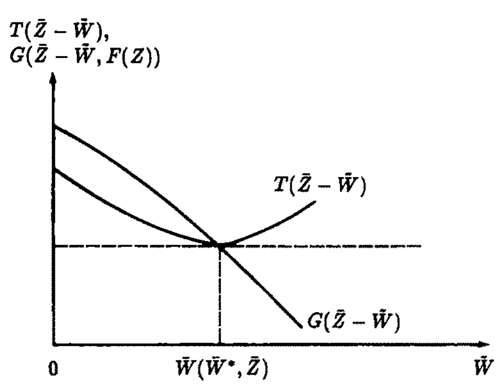


FIGURE 1

Hence, the expected gain from maintained cooperation is smaller if the amount of cooperative military expenditures at all realizations of Z is larger.

The subgame perfection constraint (13) requires that regardless of the value taken on by Z_0, the fixed, expected gain from maintained cooperation, (18), exceed the temptation to deviate from $\tilde{W}(Z_0)$, given by the difference between (15) and (17). The necessary and sufficient condition for threats and punishments to serve as a perfect substitute for the ability to make binding commitments is that for all possible realizations of $Z_0, T(Z_0 - 0) - G(Z - 0, F(Z)) \leq 0$.

In the special case that $Z_t = \bar{Z}$ for $t = 0, 2, \ldots, \infty$, threats and punishments support the disarmament solution provided that the discount rate is not too large. In this fixed, nonstochastic environment, the cooperative policy, $\tilde{W}$ is independent of time. For this special case, the relationship between the temptation and the gain, as functions of $\tilde{W}$, is drawn in Figure 1. As shown in the figure, the gain equals the temptation at $\tilde{W} = (\overline{W}(\overline{W}^*, \bar{Z}))$. Moreover, the slope of $G((\bar{Z} - \tilde{W}, \bar{Z}))$ is strictly greater in magnitude than that of $T(\bar{Z} - \tilde{W})$ for $0 \leq \tilde{W} < \overline{W}(\overline{W}^*)$ provided that $\beta \geq (1 - \alpha)/(1 + \beta)$.[22] At $\tilde{W} = 0$, then, the gain from maintained cooperation exceeds the temptation for opportunistic behavior. Since the disarmament policy satisfies the subgame perfection constraint (13), it is the cooperative equilibrium. In other words, the folk theorem of game theory holds in the nonstochastic version of the model. If individuals do not discount the utility of future consumption too heavily, the disarmament outcome can emerge in the equilibrium of an infinite horizon game with threats and punishments (or equivalently, in the cooperative equilibrium).

More generally, however, the disarmament solution is not necessarily supported in a cooperative equilibrium for all values of the endowment. If the distribution of Z is unbounded, then threats and punishments—including punishments more severe than the trigger strategy—cannot support the disarmament outcome. Even if the distribution is bounded, where $[Z_l, Z_h]$ is the support of Z, there is nothing to guarantee that the disarmament policy is consistent with the subgame perfection constraint, (13), —that is, $T(Z_0 - 0) - G(Z - 0, F(Z)) \leq 0$—for all possible values of Z_0. Nevertheless, the folk theorem holds in the sense that for a given bounded distribution, there exists a $\beta^* < 1$ such that $\beta \geq \beta^*$ is a sufficient condition for disarmament to be the cooperative equilibrium, supported by a trigger strategy, for all t. In the stochastic environment, the critical value of the discount factor, β^*, depends on the distribution of the endowment, and will generally be larger than $(1 - \alpha)/(1 + \beta)$.[23]

B. *Military Spending in the Cooperative Equilibrium*

When a threat of a reversion to the opportunistic policy is not sufficient to support the efficient outcome for all possible realizations of Z, military spending with such a threat can deter opportunistic behavior. To characterize military spending in the cooperative equilibrium, suppose that $\tilde{W}(Z_t) = 0$—that is, disarmament—does not satisfy the subgame perfection constraint (13) for $Z_t > \phi$. For values of the endowment in excess of ϕ, the optimal contingent cooperative policy, $\tilde{W}(Z_t)$, which specifies the lowest incentive compatible amounts of military spending for each realization of Z_t, satisfies (13) with an equality. Treating (13) as an equality and using the implicit function theorem, one can establish that

---

[22] If, for example, $\alpha = 1/2$, a discount factor greater than or equal to 0.367 would satisfy this condition.

[23] Note that for more severe punishments, β^* would be smaller.

$$\frac{\partial \tilde{W}(Z_t)}{\partial Z_t} = -\frac{\partial T\big(Z_t - \tilde{W}(Z_t)\big)/\partial Z_t}{\partial T\big(Z_t - \tilde{W}(Z_t)\big)/\partial \tilde{W}(Z_t)}$$
$$= 1.$$

The cooperative policy involves larger amounts of military spending for larger values of the current realization of the endowment. To offset the larger temptation for larger values of Z_t, it is necessary to specify larger $\tilde{W}$ for those possible values of Z_t. In fact, since

$$\frac{\partial T\big(Z - W(\tilde{Z})\big)}{\partial Z} = -\frac{\partial T\big(Z - W(\tilde{Z})\big)}{\partial W(Z)}$$

(as mentioned above), $\tilde{W}^*(Z_t)$ increases one for one with increases in Z_t in excess of ϕ. Thus, the contingent cooperative policy function, $W(\tilde{Z}_t)$, is flat for $Z_t < \phi$ and kinked at $Z_t = \phi$ with slope equal to 1:

$$(19) \quad \tilde{W}(Z_t) = \begin{cases} 0 & \text{if } Z_l \leq Z_t \leq \phi; \\ Z_t - \phi & \text{if } \phi < Z_t \leq Z_h, \end{cases}$$

where $k \leq \phi < Z_h$. Note that ϕ might be less than Z_l, the lowest possible realization of Z_t. In this case, $\phi \varepsilon [k, Z_l)$, and there is positive military spending for all values of the endowment. In any case, military spending by one nation serves to deter the opponent from deviating from the cooperative policy to extract tribute.

According to equation (19), if $\phi = k$, then the cooperative military spending is identical to military spending in the opportunistic equilibrium—that is $\tilde{W}^*(Z_t) = \overline{W}(\overline{W}^*, Z_t)$. However, if $\phi > k$, the level of military spending in the cooperative equilibrium will be less than in the opportunistic equilibrium for all Z. Furthermore, if $\phi > Z_l$, then there are some realizations of the endowment for which the optimal contingent cooperative policy involves disarmament.

Because ϕ is the largest realization of the endowment that would be consistent with disarmament, ϕ is the largest number that satisfies

$$(20) \quad T(\phi - 0) = G\big(Z - \tilde{W}(Z), F(Z)\big),$$

where $\tilde{W}(Z)$ is given by equation (19). From equations (15) and (17),

$$T(\phi - 0)$$
$$= \left[(1+\beta)\ln\left(\phi + \frac{1-\alpha}{\alpha}\left(\frac{\alpha}{a}A\right)^{\frac{1}{1-\alpha}}\right) \times \frac{1}{1+\beta} + \beta \ln \beta a\right]$$
$$- \left[(1+\alpha\beta)\ln\frac{\phi}{1+\alpha\beta} + \beta \ln A + \alpha\beta \ln \alpha\beta\right].$$

And, from equations (18) and (19),

$$G\big(Z - \tilde{W}(Z), F(Z)\big)$$
$$= \begin{cases} \text{If } \dfrac{\beta^2}{1-\beta^2}(1+\alpha\beta)\left[\displaystyle\int_{Z_l}^{\phi} \ln\frac{Z_t}{k}F(Z) + (1-F(\phi))\ln\frac{\phi}{k}\right] & \text{if } Z_l \leq \phi \leq Z_h; \\ \text{If } \dfrac{\beta^2}{1-\beta^2}(1+\alpha\beta)\ln\dfrac{\phi}{k} & \text{if } k \leq \phi < Z_l. \end{cases}$$

The value of ϕ that satisfies (20) will be the constant difference between the endowment and the level of cooperative military spending when military spending is positive—that is, according to (19), $\phi = Z_t - \tilde{W}(Z_t)$ for $Z_t > \phi$.

Assume that the necessary and sufficient condition for disarmament to prevail for all t, given by $T(Z_h - 0) \leq G(Z - 0, F(Z))$ (or equivalently, $\phi = Z_h$), is not met. Then the value of ϕ that satisfies (20) is an element of $[k, Z_h)$. It can be shown that if the rate of time preference satisfies the condition $\beta \geq (1-\alpha)/(1+\beta)$, then the value of ϕ that satisfies (20) will be greater than Z_l. To see that $\beta \geq (1-\alpha)/(1+\beta)$ implies that $\phi > Z_l$, observe that if $Z - \tilde{W}(Z)$ equals k for all Z, then $T(\phi - 0) = G(Z - \tilde{W}(Z), F(Z))$. Provided that $\beta \geq (1-\alpha)/(1+\beta)$, the difference $T(\phi - 0) - G(Z - \tilde{W}(Z), F(Z))$ decreases with increases in the constant $Z - \tilde{W}(Z)$ for $Z > \phi$, over the interval $[k, Z_l]$.[24] Hence, if the rate of time preference is not too large, there will be some values of the endowment for which threats and punishments can support the disarmament solution.

Maintaining the assumption that $\beta \geq (1-\alpha)/(1+\beta)$, so that $\phi > Z_l$, it is possible to show that the larger is the variance of the endowment, σ^2, for a given expected value of the endowment, $\bar{Z}$, the smaller is ϕ—that is,

$$\frac{\partial\phi}{\partial\sigma^2} = -\frac{\partial T(\phi-0)/\partial\sigma^2 - \partial G(Z-\tilde{W}(Z), F(Z))/\partial\sigma^2}{\partial T(\phi-0)/\partial\phi - \partial G(Z-\tilde{W}(Z))/\partial\phi} < 0.$$

Because the greatest number that satisfies (21), ϕ, is less than Z_h,—that is, because $T(Z_h - 0) - G(Z - 0, F(Z)) > 0$—the denominator of the above expression is positive. To see that the numerator of the above expression is nonnegative, note that $\partial T(\phi - 0)/\partial\sigma^2$ equals zero, since the temptation is independent of the variance of the endowment for any given ϕ. Using theorems of second-order stochastic dominance, it is easy to verify that $\partial G(Z - \tilde{W}(Z), F(Z))/\partial\sigma^2$ is nonpositive.[25] The proposition follows from the assumptions that the marginal product of the peaceful investment is diminishing and that the marginal product of war goods is constant, which, in conjunction with the assumption of diminishing marginal utility, imply that the expected gain from maintained cooperation is increasing in the endowment at a diminishing rate.

Similarly, it is possible to show that a smaller expected value of the endowment, $\bar{Z}$, implies a smaller ϕ, for a given σ^2—that is,

$$\frac{\partial\phi}{\partial\bar{Z}} = -\frac{\partial T(\phi-0)/\partial\bar{Z} - \partial G(Z-\tilde{W}(Z), F(Z))/\partial\bar{Z}}{\partial T(\phi-0)/\partial\phi - \partial G(Z-\tilde{W}(Z), F(Z))/\partial\phi} > 0.$$

Again, the denominator of the above expression is positive. Notice that the temptation is independent of the expected value of the endowment for any given ϕ. Hence, $\partial T(\phi - 0)/\partial\bar{Z}$ equals zero. Upon inspection of (20), where the expected gain from maintained cooperation is defined for $\phi > Z_l$, it becomes obvious that $\partial G(Z - \tilde{W}(Z), F(Z))/\partial\bar{Z}$ is positive. The intuition here is as follows. Holding ϕ fixed and increasing the expected value of the endowment means that conditional on $Z < \phi$, when there is disarmament and the two-period expected utility from cooperation depends on Z, the two-period expected utility from cooperation is higher. And, conditional on $Z > \phi$, the two-period expected utility from cooperation is unchanged. Since the utility obtained in the punishment is independent of Z, the expected gain from maintained cooperation increases with increases in $\bar{Z}$. Hence, the numerator is negtative and the whole expression is positive.

From equation (19), a smaller ϕ implies that for a given realization of Z_t, the level of cooperative military spending is higher.[26] Therefore, if the cooperative equilibrium is empirically relevant, assuming that the cooperative equilibrium threshold level of the endowment, ϕ, is greater than the lowest possible realization of the endowment, Z_l, one would find, for a given realization of the endowment, that the associated level of military spending is larger, the larger is the variance of the endowment and the smaller is the mean value of the endowment. Alternatively, if the opportunistic equilibrium were empirically relevant, one would find no relationship between a nation's military spending and the variability or the mean value of the endowment, for a given realization of the endowment.

---

[24] See Figure 2 where the difference $T(\phi-0) - G(Z - W(Z), F(Z))$ is drawn as a function of $Z - \tilde{W}(Z)$ for $Z > \phi$, assuming that $\beta \geq (1-\alpha)/(1+\beta)$.

[25] See the Appendix for a proof.

[26] Also, see Figure 3.

It is important to note that while these empirical implications do not depend on the specifications made for utility and the peaceful investment technology, (5), they do depend on two assumptions. First, these implications require that the threshold level of the endowment is greater than the lowest possible realization of the endowment, such that there are some values of the endowment for which disarmament prevails. The assumption is important for the proofs sketched above. If instead the equilibrium cooperative policy always involves positive military spending, then the expected gain will be independent of the distribution of the endowment. Even so, it is still possible to make some positive predictions. Specifically, although the hypothetical changes in the distribution of Z would not have an impact on the difference $T(\cdot)-G(\cdot)$ over the range $[k, Z_l]$, these changes would alter the difference in the interval $(Z_l, Z_h]$. An increase in $\bar{Z}$ or a decrease in σ^2 would push $T(\cdot)-G(\cdot)$ down over the interval $(Z_l, Z_h]$. Since ϕ is the largest number that makes $T(\cdot)$ equal to $G(\cdot)$, if the effect is large enough to result in an intersection of the difference $T(\cdot)-G(\cdot)$ and the zero horizontal axis (see Figure 2) over the interval $(Z_l, Z_h]$, then, assuming that the original ϕ is less than Z_l, the change in the distribution would imply a discontinous jump in $\phi \varepsilon [Z_l, Z_h]$, or equivalently, a drop in cooperative military spending for each realization of the endowment.

Second, and more importantly, these implications depend on the linear specification of the tribute function. If instead the marginal product of war goods were diminishing, then the expected gain could be independent of the endowment or decreasing in the endowment at an increasing rate. As a consequence, exercises involving hypothetical changes in the distribution might not yield implications that would enable one to distinguish empirically the opportunistic equilibrium from the cooperative equilibrium.[27]

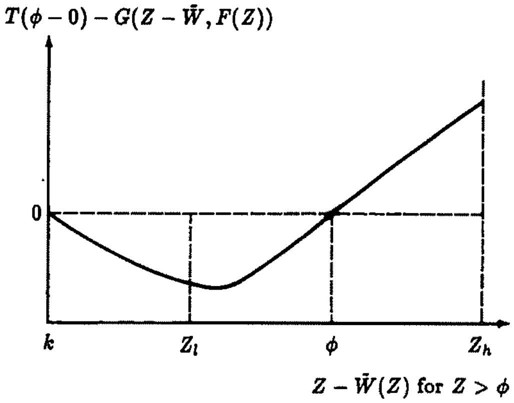
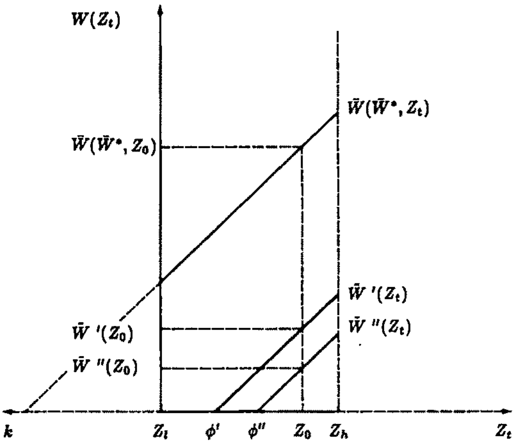


FIGURE 2



$\phi' < \phi'' \rightarrow \bar{W}'(Z_t) > \bar{W}''(Z_t)$

FIGURE 3

Generally, in a stochastic environment, threats and punishments do not necessarily serve as a perfect substitute for the ability to make commitments. Although the punishment strategy employed in the analysis does not involve the most severe punishment, consideration of more severe punishments will not qualitatively alter the discussion above. Specifically, the most severe punishment can only yield a fixed expected gain to maintained cooperation. The lowest utility that a government could expect to obtain if it followed its mimimax strategy and the opponent devoted all of its resources to military spending starting in period $t = 2$ establishes an upper bound to the gain from maintained cooperation that could be expected by the governments with the most severe punishment strategy. There is nothing to guarantee that the fixed, expected gain will dominate the temptation, for all realizations of the endowment, where the cooperative policy is the disarmament policy. While strategic threats and punishments would be more likely to support the efficient solution if the discount rate is low, a low discount rate is not always sufficient to ensure that the efficient solution emerges in a cooperative equilibrium. Hence, while threats and punishments can reduce the magnitude of the resources wasted away in military spending, they cannot necessarily eliminate the time-inconsistency problem of the commitment to peaceful policies.

---

[27] Deriving and implementing more general testable empirical implications of the model are left for future research.

III. Conclusion

This paper has investigated the role of strategic considerations in an economic model of international conflict. The analysis has shown that when there is repeated interaction between nations, threats and punishments, which at least serve as a partial substitute for a government's ability to make commitments, might be sufficient for sustaining the *ex ante* optimal, but otherwise time-inconsistent, disarmament policy. Threats and punishments are more likely to support the commitment policy, as the discount factor approaches one.

If the fixed expected gain from maintained cooperation does not exceed the temptation to deviate from the peaceful policy for all possible realizations of the endowment, then the model predicts that there will be some periods of positive production of armaments. Because the temptation to deviate from the cooperative policy is increasing in the endowment, for values of the endowment that fall below an endogenously determined threshold value, the contingent cooperative policy involves only peaceful policies, whereas for values of the endowment that exceed the threshold value, the cooperative policy involves positive military expenditures, the magnitude of which is dependent on the current realization of the endowment.[28] Hence, fluctuations in military spending are an endogenous result of fluctuations in the endowment in both the opportunistic and the cooperative equilibria. This theoretical result casts doubt on the validity of empirical work that treats military expenditures as exogenous.[29]

Although the relationship between military spending and the endowment is positive regardless of whether governments are acting cooperatively or opportunistically, there are qualitative differences in this relationship depending on whether governments are acting cooperatively or opportunistically. In particular, if governments are acting cooperatively, the larger is the variance of the endowment and smaller is the mean value of the endowment, the smaller is the equilibrium threshold level of the endowment. Consequently, for a given realization of the endowment, a nation's military spending is positively related to the variablity of the endowment and negatively related to the mean value of the endowment. In contrast, if nations are acting opportunistically, their military spending would not be related to the variance of the endowment or the mean value of the endowment, for a given realization of the endowment.

---

[28] Although the notion of positive military expenditures in a cooperative equilibrium might seem counterintuitive, this result is very similar to the results of Grossman and Van Huyck (1988), who study the role of reputation in the problem of sovereign debt, interpreted as a contingent claim. Positive military expenditures on armaments in the cooperative equilibrium would correspond to their notion of excusable default, which is to be clearly distinguished from repudiation or a reversion to the opportunistic policy.

[29] See for example Hall (1986), who assumes that military spending is exogenous to identify the role of consumption in macroeconomic fluctuations. As discussed above—fn. 1—the existing empirical evidence concerning the (statistical) causal relation between military spending and other aggregate economic variables is ambiguous.

Finally, it is important to observe that in the cooperative equilibrium, military spending is not directly driven by the possibility of extracting resources from an enemy nation. Rather, military spending serves to deter the nations from building armaments in an opportunistic effort to obtain tribute.

APPENDIX

To see that a larger variance of the endowment implies an equal or smaller expected gain from maintained cooperation, consider two distributions, $F_1(Z)$ and $F_2(Z)$, that have identical supports and mean values:

$$\int_{Z_l}^{Z_h} Z[dF_1(Z) - dF_2(Z)] = 0,$$

or equivalently,

$$\text{(A1)} \qquad \int_{Z_l}^{Z_h} [F_1(Z) - F_2(Z)]\, dZ = 0.$$

If $dF_1(Z)$ and $dF_2(Z)$ satisfy the single crossing property, there exists a $Z' \epsilon (Z_l, Z_h)$ such that

$$\text{(A2)} \quad F_1(Z) - F_2(Z) \begin{cases} \leq 0 & \text{if } Z' \leq Z \leq Z_h; \\ \geq 0 & \text{if } Z_l \leq Z \leq Z'. \end{cases}$$

The two distributions that satisfy (A1) and (A2) also satisfy the following:

$$\text{(A3)} \qquad \int_{Z_l}^{y} [F_1(Z) - F_2(Z)]\, dZ \geq 0,$$

for all $y \epsilon [Z_l, Z_h]$. The expressions in (A1) and (A3) mean that $F_2(Z)$ dominates $F_1(Z)$ in the sense of second-order stochastic dominance. $F_1(Z)$, which is obtained from a mean preserving spread of $F_2(Z)$, has a larger variance than $F_2(Z)$. (See Michael Rothschild and Joseph Stiglitz, 1970).

Using (20) for $\phi \geq Z_l$, a switch from $F_2(Z)$ to $F_1(Z)$, (an increase in the variance) results in the following change in the expected gain from maintained cooperation:

$$\text{(A4)} \qquad \int_{Z_l}^{\phi} \ln Z[dF_1(Z) - dF_2(Z)] + [F_2(\phi) - F_1(\phi)]\ln\phi.$$

Integrating by parts, (A4) becomes

$$\text{(A5)} \quad \ln Z[F_1(Z) - F_2(Z)]\Big|_{Z_l}^{\phi} - \int_{Z_l}^{\phi} \frac{1}{\phi}[F_1(Z) - F_2(Z)]\, dZ + [F_2(\phi) - F_1(\phi)]\ln\phi,$$

which simplifies to the following:

$$\text{(A6)} \qquad -\int_{Z_l}^{\phi} \frac{1}{\phi}[F_1(Z) - F_2(Z)]\, dZ,$$

since $F_1(Z_l) = F_2(Z_l) = 0$. Integrating (A6) by parts gives the following expression for the change in the expected gain due to the mean preserving spread of the distribution:

$$\text{(A7)} \qquad -\frac{1}{\phi}[\hat{F}_1(\phi) - \hat{F}_2(\phi)] + \int_{Z_l}^{\phi} -\frac{1}{\phi^2}[\hat{F}_1(Z) - \hat{F}_2(Z)],$$

where $\hat{F}(Z) \equiv \int_{Z_l}^{\phi} F(Z)\, dZ$ and $\hat{F}(Z_l) = 0$. By virtue of (A3), the expression above is nonpositive. Hence, an increase in the variance of the endowment implies an equal or smaller expected gain from maintained cooperation.

REFERENCES

Abreu, Dilip, "External Equilibria of Oligopolistic Supergames," *Journal of Economic Theory*, June 1986, *39*, 191–225.

Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.

Barro, Robert and Gordon, David, "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, July 1983, *12*, 101–21.

Brams, Steven and Kilgour, Marc, "The Path to Stable Deterrence," in U. Luterbacher and M. Ward, Boulder, eds., *Dynamic Models of International Conflict*, Boulder, CO: Lynn Rienner Publishers, 1985, 11–25.

______ **and** ______, "Deterrence Versus Defense: A Game-Theoretic Model of Star Wars," C. V. Starr Center for Applied Economics, Research Report No. 86-35, 1986.

Brams, Steven and Wittman, Donald, "Nonmyopic Equilibria in 2×2 Games," *Conflict Management and Peace Science*, Fall 1981, *6*, 39–62.

Brito, Dagobert, "A Dynamic Model of an Armaments Race," *International Economic Review*, June 1972, *13*, 359–75.

______ **and Intriligator, Michael,** "Conflict, War, and Redistribution," *American Political Science Review*, Winter 1985, *79*, 943–57.

Canzoneri, Matthew and Gray, Jo Anna, "Monetary Policy Games and the Consequences of Non-Cooperative Behavior," *International Economic Review*, October 1985, *26*, 547–64.

______ **and Henderson, Dale,** "Strategic Aspects of Macroeconomic Policymaking in Interdependent Economies: Reputation in Repeated Games," mimeo., 1986.

______ **and** ______, "Is Sovereign Policymaking Bad?" *Carnegie-Rochester Conference Series on Public Policy*, Spring 1988, *28*, 93–140.

Feinman, Joshua, Garber, Peter, and Garfinkel, Michelle, "Financial Warfare: Attacks on Credit Mechanisms," mimeo., 1986.

Friedman, James, "A Non-cooperative Equilibrium for Supergames," *Review of Economic Studies*, January 1971, *38*, 1–12.

Garcia-Mila, Teresa, "Government Purchases and Real Output: An Empirical Analysis and Equilibrium Model with Public Capital," mimeo., 1987.

Grossman, Herschel and Van Huyck, John, "Seingniorage, Inflation, and Reputation," *Journal of Monetary Economics*, July 1986, *18*, 21–34.

______ **and** ______, "Sovereign Debt as a Contingent Claim: Excusable Default, Repudiation, and Reputation," *American Economic Review*, December 1988, *78*, 1088–97.

Hall, Robert, "The Role of Consumption in Economic Fluctuations," in *The American Business Cycle*, Robert Gordon, ed., Chicago: University of Chicago Press, 1986, 237–55.

Intriligator, Michael, "Strategic Considerations in the Richardson Model of Arms Races," *Journal of Political Economy*, April 1975, *83*, 339–53.

______, "Research on Conflict Theory: Analytic Approaches and Areas of Application," *Journal of Conflict Resolution*, June 1982, *26*, 307–27.

______ **and Brito, Dagobert,** "Can Arms Races Lead to the Outbreak War?," *Journal of Conflict Resolution*, March 1984 *28*, 63–84.

Maoz, Zeev, "Decision-Theoretic and Game-Theoretic Models on International Conflict," in *Dynamic Models of International Conflict*, U. Luterbacher and M. Ward, eds., Boulder, CO: Lynn Rienner Publishers, 1985, 76–11.

Milward, Alan, *The New Order and the French Economy*, Oxford: Clarendon Press, 1970.

North, Douglas, *Structure and Change in Economic History*, New York: W. W. Norton & Co., 1981.

Olson, Mancur, Jr. and Zeckhauser, Richard, "An Economic Model of Alliances," *Review of Economics and Statistics*, August 1966, *48*, 266–79.

Riddell, Tom, "U.S. Military Power, the Terms of Trade, and the Profit Rate," *American Economic Review Papers and Proceedings*, May 1988, *78*, 60–65.

Robbins, Lionel, *The Economic Causes of War*, New York: Macmillan, 1940.

Rothschild, Michael and Stiglitz, Joseph, "Increasing Risk I: A Definition," *Journal of Economic Theory*, September 1970, *2*, 225–43.

Schelling, Thomas, *The Strategy of Conflict*, Cambridge, MA: Harvard University Press, 1960.

Shepherd, William, "Self Interest and National Security," *American Economic Review Papers and Proceedings*, May 1988, *78*, 50–54.

Shubik, Martin, "Game Theory Models of Strategic Behavior and Nuclear Deterrence,"Cowles Foundation Discussion Paper No. 829, 1987.

Smith, R. P., "Alternative Models of Military Expenditure," Birbeck College Discussion Paper in Economics, No. 87-17, 1987.

van der Ploeg, F. and de Zeeuw, A. J., (1986a) "Conflict Over Arms Accumulation in Market and Command Economies," mimeo., 1986.

______ **and** ______, (1986b) "Perfect Equilibrium in a Model of Competitive Arms Accumulation," mimeo., 1986.

Zagare, Frank, "The Pathologies of Unilateral Deterrence," in *Dynamic Models of International Relations*, U. Luterbacher and M. Ward, eds., Boulder, CO: Lynn Rienner Publishers, 1985, 52–75.

Stackelberg and Marshall

By ARTHUR J. ROBSON*

This paper advocates a generalized N-firm Stackelberg model as a plausible testable alternative description of oligopoly. A pure-strategy equilibrium must exist for this model. The main result is that efficiency obtains in the limit as the scale of each firm is shrunk relative to demand. This is demonstrated for the case of U-shaped average cost and also for that of natural monopoly. (JEL 611)

The standard classroom model of oligopoly is, of course, that due to A. A. Cournot. In its most familiar form, there is linear demand, zero production cost, and an exogenously given number of firms. As this number of firms is increased, it is reassuring to find that the efficient, competitive level of industry output is approached. This limiting property serves to check the validity of the Cournot model, given the perhaps greater direct intuitive appeal of competition. From another viewpoint, this property shows that competition can be viewed as the limiting form of a game with simultaneous quantity decisions taken by all firms.

It is important then to generalize the above very special case. A difficulty which is encountered immediately with the introduction of arbitrary downward-sloping demand is that a Cournot equilibrium need not exist. (See Maurice McManus, 1964, or John Roberts and Hugo Sonnenschein, 1977). The problem is just that reaction functions can now jump and need not intersect as required for a Cournot equilibrium. William Novshek (1980), on the other hand, presents a model of Cournot equilibrium with general demand, identical U-shaped average cost curves, and free entry and exit. Novshek is then able to show that a Cournot equilibrium must *eventually* exist, as the scale of each firm is shrunk relative to demand. Furthermore, as the shrinkage continues, the number of active firms tends to infinity, and the Cournot equilibria converge to the competitive equilibrium.

The present paper investigates the properties of a generalized Stackelberg equilibrium. It is assumed, that is, that firms choose quantities in a strict sequence, not simultaneously, and that each firm is fully informed of previous choices. This Stackelberg model was originally described with just two firms. (See H. von Stackelberg, 1934, but also F. Y. Edgeworth, 1925). The following conceptual objection was first raised by Edgeworth. How is it determined which of two possibly identical firms is to be anointed as the leader and which relegated to the role of follower? To put this another way: There are now two Stackelberg equilibria with claims on our attention. If the Cournot equilibrium were always unique, this would certainly be a telling point in its favor. However, although the above simple pedagogic example has a unique Cournot equilibrium, uniqueness will generally fail to hold as soon as the model is generalized. In particular, free entry is likely to lead to nonuniqueness. For example, suppose that demand and average cost are such that monopoly output by any one firm suffices to deter any other firm from producing. This is a simple kind of Cournot equilibrium with free entry. Note, however, that now there are as many such equilibria as firms since any one of the firms can be chosen as the solitary active firm. Indeed, a Cournot equilibrium with free entry does not generally even determine a unique number of active firms. Thus Edgeworth's criticism of

---

*Professor of Economics, University of Western Ontario, London, Canada. I acknowledge research funding from the Social Sciences and Humanities Research Council of Canada. I also thank the referees of this *Review* for their comments and Tom Rutherford for computer programming assistance.

what later come to be called Stackelberg equilibrium applies with similar force to a Cournot equilibrium with free entry.

Nonetheless, if an alternative asymmetric Stackelberg equilibrium is to be a plausible snapshot description of some industry at a given time, there should be an historical source of differentiation among the firms. Without this, it seems hard to dispute that an explicitly symmetric equilibrium concept remains more appropriate. A natural source of historical differentiation is sunk cost of capacity. Indeed, Michael Spence (1977) demonstrates how such costs lead to Stackelberg—like equilibria, given an exogenous sequence for capacity choices. This argument in favor of Stackelberg at the expense of Cournot can be made more forcefully. That is, even if firms *endogenously* select the dates of their capacity choices, and simultaneous choices are not ruled out, asymmetric timing is likely to be the equilibrium outcome. (Drew Fudenberg and Jean Tirole, 1985, make a similar claim). To focus just on this latter point, consider the following simple model of duopoly in which capacity and quantity choices are locked together, and this combined choice can be made just once.

The market for the good exists only at time $t=0$ with price determined as $P=2-(q_1+q_2)$, where q_1 and q_2 are the quantities produced by firms 1 and 2, respectively. Each firm is limited to quantities from the set $\{0,1/4,1/3,1/2\}$ but can make this choice either at time $t=-1$ with unit cost $(1+r)$; or at time $t=0$ with unit cost 1. It is assumed $r>0$, so early choice is costly in this sense. The interval of time involved here is taken to be short, and hence r is also small. Now it is not hard to see that the only pure-strategy "subgame-perfect equilibria" of this duopoly game are of Stackelberg type. That is, either 1 chooses $1/2$ at $t=-1$ and 2 chooses $1/4$ at $t=0$, or vice versa. Why, for example, is it not an equilibrium for both firms to choose at $t=-1$? Whatever choice firm 1, say, makes at $t=-1$, it would be better to make this choice at $t=0$ instead, since firm 2 could not react to this deviation. Why, again, is it not an equilibrium for both firms to choose $t=0$? In this case, the force of subgame perfectness is just to require that the Cournot equilibrium would then be played, with each firm playing $1/3$, and obtaining thereby a profit of $1/9$. If, however, firm 1, for example, deviates from this supposed equilibrium by choosing $1/2$ at $t=-1$ instead, firm 2 can only react by playing the best-reply to this, namely, $1/4$. (This is, again, dictated by subgame perfectness). But now firm 1 has a profit of $(1/4-r)/2$ which exceeds $1/9$ if r is small enough.

A more general example is presented in the appendix, as Example A. This models precommitment explicitly in terms of irreversible capacity investment. Each of the two firms is now free to select any amount of capacity at $t=-1$, any amount of capacity at $t=0$, and any amount of output not exceeding total capacity, also at $t=0$. Again, however, when the interest rate is small, the only subgame perfect equilibria in pure strategies owe more to Stackelberg than to Cournot.

In the preceding discussion, a case was made that Stackelberg equilibria compare favorably with Cournot, from a conceptual viewpoint. Another salient testable model of oligopoly is that of "contestability." (See William Baumol, John Panzar, and Robert Willig, 1982, or, for an overview, Baumol, 1982). It is useful then to also compare the assumptions of the present model with those of contestability. One key difference is that sunk costs are required to be present here; they are required to be absent in the contestability model. (See Marius Schwartz and Robert Reynolds, 1983). Another key difference is the type of strategies used by the firms. In the contestability model, it is assumed that incumbent firms' *prices* would be fixed, at least briefly, after entry, if entry occurs. In the present model, firms fix *outputs* (perhaps via capacities) in a strict sequence. (William Brock and José Scheinkman, 1983, also build a model in which sunk capacities determine outputs, but which otherwise remains similar in spirit to the contestability model).

The first difference in results concerns the existence of an equilibrium. It is noted here that recent results in game theory guarantee existence of a generalized Stackelberg equilibrium in pure strategies under mild conditions. As noted above, the same cannot be said of Cournot equilibrium. Further, it is easy to find examples where no sustainable equilibrium exists in the contestability model. (See Baumol et al., 1982.)

Perhaps the best way of viewing the differences in other results of the three models is in terms of efficiency. Contestability leads to constrained efficiency, but this is not true here nor in the Cournot model, with free entry. Such efficiency results as do apply in the Cournot case with free entry are asymptotic in nature, when the scale of each firm is shrunk relative to demand, as noted above. It is shown that similar results apply for the Stackelberg model with free entry.

Section I considers, then, a situation in which identical firms produce a homogeneous good, each firm having U-shaped average cost. There is an arbitrary downward-sloping demand. As the scale of each firm is shrunk, the Stackelberg equilibria with free entry are shown to converge to the competitive equilibrium. This asymptotic efficiency result demonstrates that competition might be viewed as the limiting form of a game with sequential quantity decisions. When added to the previous results on the convergence of Cournot equilibria with free entry, it proves that competition is then, in this sense, robust to the specification of strategic timing.

Section II considers the less general but still empirically relevant case in which average cost declines always. Each firm, that is, is an identical "natural monopoly." (Note that Baumol et al., 1982, use this term in a strictly more general sense.) This also yields an interesting test case for a comparison of the three models.

Recall, first, that Martin Weitzman (1983) has argued that natural monopoly is inconsistent with contestability. That is, since there can be no sunk costs, fixed costs can be "convexified" by a suitable on-off regime over time. This argument does not apply to the present model just because sunk costs are present, are indeed an essential feature. Baumol et al. (1983) do not fully accept Weitzman's viewpoint. It is still of interest to compare the predictions of the present model with those made on the basis of contestability.

The Stackelberg equilibrium with free entry is readily explicitly constructed here. As is true with contestability, but is not true for the Cournot equilibrium with free entry, there is a *single* active firm. As is not true with contestability, but is true for this Cournot equilibrium, there are usually positive profits. In a Stackelberg equilibrium with free entry the first active firm merely renders further firms inactive. Clearly then this Stackelberg equilibrium is usually less socially efficient than is the contestability outcome. However, an example demonstrates that the Stackelberg equilibrium is likely to be more efficient than Cournot, given free entry, just because a single firm is, per se, desirable with natural monopoly. Finally it is shown that the Stackelberg equilibrium with free entry predicts a fully efficient level of output in the limit as the scale of each firm is shrunk relative to demand, due to the increasing pressure of *potential* competition. (This presumably would be true, as well, for the other two models.) This is true here despite the existence always of a single active firm.

I. Stackelberg Oligopoly—*U*-Shaped Average Cost

Consider an industry with a demand and cost structure based on that for classroom discussions of competition. To be precise:

ASSUMPTION 1: *Cost. Each firm has average cost given by the continuously differentiable function*

$$a\colon (0,\infty) \to (0,\infty)$$

$$\begin{aligned} a'(q) &< 0 \quad \forall q \in (0, q^*) \\ &= 0 \quad \text{iff} \quad q = q^* \\ &> 0 \quad \forall q > q^* \end{aligned}$$

for some $q^* > 0$. *The level of minimum aver-*

age cost is denoted

$$P^* = a(q^*).$$

Total cost is then

$$\begin{aligned} c(q) &= 0 \qquad q = 0 \\ &= q \cdot a(q) \quad \forall q > 0, \end{aligned}$$

which is certainly continuous $\forall q > 0$ *but need not be continuous at* $q = 0$.

ASSUMPTION 2: *Demand. Inverse demand is a continuous function*

$$P\colon [0,\infty) \to [0,\infty),$$

which is continuously differentiable, with

$$P'(Q) < 0 \text{ whenever } P(Q) > 0.$$

In addition, it is assumed that

$$\lim_{Q \to \infty} P(Q) < P^* < P(0) < \lim_{q \to 0} a(q),$$

where P^* *and* $a(q)$ *are defined as in Assumption* 1. *The first two inequalities are needed in order that demand and the limiting competitive supply at* P^* *intersect nontrivially. The third inequality rules out the possibility that a monopoly might find "infinitesimal" output to be most profitable. It is used here to obtain a lower bound on output per firm and hence an upper bound on the number of active firms. It also implies that the possible discontinuity in the cost function at zero causes no difficulty in proving existence of an equilibrium.*

The unique solution of

$$P(Q) = P^*$$

is denoted Q^*. *In addition the unique solution of*

$$a(q) = P(0), q < q^*$$

is denoted $\bar{q} > 0$. *This is the required lower bound on an active firm's output.*

Definition 1: Scaled Down Cost. The model here is to be parameterized by a scale factor $\alpha \in (0,1]$, which transforms each firm's average and total cost according to

$$a(\alpha,q) = a\left(\frac{q}{\alpha}\right), c(\alpha,q) = \alpha \cdot c\left(\frac{q}{\alpha}\right).$$

This is as in Novshek (1980). This implies that $a(\alpha,q)$ remains *U*-shaped, with the same level of minimum average cost, P^*, attained now at output αq^*. The lower bound on each active firm's output becomes $\alpha\bar{q}$.

Definition 2: Stackelberg Oligopoly Game. Specification of any game requires that the *strategies* of all players be defined as well as the resulting *payoffs*.

Consider then N firms with identical costs $c(\alpha,q)$ as above. They make sequential choices of outputs, firm 1 first, firm N last. It involves no loss of generality to require each firm to choose a quantity from the set

$$C = \{0\} \cup [\alpha\bar{q}, Q^*],$$

because choice of a positive quantity less than $\alpha\bar{q}$ or one greater than Q^* guarantees a strict loss, regardless of other firm's choices. On the other hand, choice of zero output entails zero profits and no firm need do worse than this. (Mathematically, C is closed, bounded, and hence compact, even though it is not connected.)

In order to completely specify firm 1's behavior, it is sufficient to indicate merely a choice of quantity from C. Hence firm 1's strategy set is

$$S_1 = C.$$

However, complete description of firm i's behavior $i = 2,\dots,N$ must indicate a choice of quantity from C as a *function* of the vector of prior quantity choices by firms $1,\dots,i-1$. (Recall that firm i is informed of these prior choices). Thus a strategy for firm $i, i = 2,\dots,N$ maps C^{i-1} into C, and firm i's strategy set is

$$S_i = \{f_i : C^{i-1} \to C\}\, i = 2,\dots,N.$$

Given a complete set of strategies for all firms $(q_1, f_2,\dots,f_N)$ a complete set of quantity choices is uniquely determined recursively as

$$q_1, q_2 = f_2(q_1), q_3 = f_3(q_1, q_2), \ldots,$$
$$q_N = f_N(q_1, \ldots, q_{N-1}).$$

The payoffs are given directly in terms of quantity vectors, as

$$\Pi_i: C^N \to R$$

$$\Pi_i(q_1, \ldots, q_N) = q_i \cdot P\left(\sum_{j=1}^{N} q_j\right) - c(\alpha, q_i),$$

where $i = 1, \ldots, N$. Assumptions 1 and 2 and Definition 1 define the functions $P(\cdot)$ and $c(\cdot, \cdot)$. (In the next section, Assumption 3 will replace Assumption 1 here. These payoffs are, in either case, continuous functions on C^N, despite or rather because C is not connected.) Given the previous paragraph, it follows that payoffs are also uniquely determined for an arbitrary *strategy* vector, which completes the basic description of the game.

Definition 3: Subgame-Perfect Equilibrium. A little informally, this is as follows. (See Reinhard Selten, 1975, for a formal definition). First, it is required that a vector of *strategies* (not *quantities*) be a Nash equilibrium. That is, it should not be possible for any firm to improve profits by a unilateral change in strategy. Even in the simplest two-firm Stackelberg model, however, recall that more than this is assumed. For the second firm is required to optimize *given* the quantity chosen by the first, whatever this might be. That is, the usual Stackelberg model disallows "empty" threats by the second firm such as driving the price to zero whenever the first firm produces anything. A "subgame-perfect equilibrium," SPE, is simply the natural extension of this very plausible idea to N firms. A "subgame" is as follows. Take an arbitrary initial sequence of quantities chosen by firms $1, \ldots, i-1$ for any $i = 2, \ldots, N$. This defines a "residual" game to be played by firms $i, \ldots, N$, where these firm's payoffs are their original payoffs, with the initial sequence of quantities substituted in everywhere. A vector of strategies is then a subgame-perfect equilibrium, SPE, if and only if it yields a Nash equilibrium not only for the entire game but also for all such subgames.

The first issue is the obvious technical question: does an SPE exist here?

THEOREM 1: *Existence of Equilibrium. Any Stackelberg oligopoly game, as in Assumptions* 1 *and* 2 *and Definitions* 1 *and* 2, *has a subgame-perfect equilibrium in pure strategies, as in Definitions* 2 *and* 3. *Furthermore, the use of a finite grid of quantities in place of C can be used to derive an approximation to this SPE.*

PROOF:

See, for example, Martin Hellwig, Wolfgang Leininger, Philip Reny, and Arthur Robson (1989) who discuss the nature of the technical difficulties involved. Pure strategies are, of course, less of a strain on one's credulity than are mixed strategies. Example B of the Appendix finds two such approximations of SPEs using a computer. □

The proof that the Stackelberg equilibria with free entry converge to the competitive equilibrium is facilitated by the following four lemmas, with the main result stated as Theorem 2. The first lemma implies that it is sufficient always to consider only a finite number of potential firms:

LEMMA 1: *Maximum Number of Active firms. Consider any subgame-perfect equilibrium of the Stackelberg oligopoly game as in Assumptions* 1 *and* 2 *and Definitions* 1, 2, *and* 3. *There exists an upper bound,* $N(\alpha)$, *say, on the number of active firms. (This notation emphasizes that this bound depends on* α. *Indeed* $N(\alpha) \to \infty$, *as* $\alpha \to 0$. *"Active" means, of course, producing strictly positive output.)*

PROOF:

An upper bound on industry output is Q^*, where $P^* = P(Q^*)$. It would be irrational to be the first firm to cause industry output to exceed Q^*. A lower bound on the output per active firm is $\alpha\bar{q}$, since any positive amount less than this guarantees a loss. An upper bound on the number of active firms is then

$$N(\alpha) = \left|\frac{Q^*}{\alpha\bar{q}}\right|,$$

where the vertical bars denote the integer part of the number enclosed. □

The next lemma shows that the competitive price level can *only* be attained as a limit:

LEMMA 2: *Marshallian Price Exceeded. Consider any Stackelberg oligopoly game as in Assumptions* 1 *and* 2 *and Definitions* 1 *and* 2, *or any subgame defined by choices* $q_1,\dots,q_{i-1}$ *for any* $i = 2,\dots,N$, *as explained in Definition* 3. *Suppose that*

$$P(Q_{i-1}) > P^*, \text{ where } Q_{i-1} = \sum_{j=1}^{i-1} q_j$$

for any of the subgames, and $Q_0 = 0$ *for the entire game. It follows that*

$$P(Q_N) > P^*$$

for any subgame-perfect equilibrium strategies, as in Definition 3.

This result simply says that if the price starts above P^*, *then equilibrium behavior from that point onward ensures it remains above* P^*.

PROOF:

As in the proof of Lemma 1, no firm will drive the price strictly under P^*, because this guarantees it a loss. All that remains then is to rule out $P(Q_N) = P^*$. Suppose indeed $P(Q_N) = P^*$ and the last active firm is firm ℓ, $\ell \geq i$. Firm ℓ must choose P^*, but the only output then consistent with nonnegative profit is $q_\ell = \alpha q^*$. It must be true then that

$$P(Q_{\ell-1} + \alpha q^*) = P^*,$$

so residual demand left after the previous $(\ell - 1)$ firms passes through the minimum point of firm ℓ's average cost curve. This seems unlikely, but such a choice of $q_\ell = \alpha q^*$ is, in any case, not optimal for firm ℓ. Note indeed that this choice of q_ℓ *strictly* deters further activity, and hence a small reduction in q_ℓ will still do this. But a small enough reduction in q_ℓ must also yield firm ℓ a strictly positive profit, given that residual demand is strictly decreasing and passes through the minimum of firm ℓ's average cost. Hence

$$P(Q_N) > P^*. \qquad \square$$

Lemma 2 is needed in the proof of the next lemma, which has greater independent interest:

LEMMA 3: *A Sufficient Condition for Activity. Consider any Stackelberg oligopoly game in Assumptions* 1 *and* 2 *and Definitions* 1, 2, *and* 3. *Define the following set for any total output level* $Q \geq 0$:

$$A(Q) = \left\{q \mid P(Q+q) > a\left(\frac{q}{\alpha}\right)\right\}.$$

That is, $A(Q)$ *represents the choices of output for any firm which yield strictly positive profits, given a previous total output of* Q, *and assuming no subsequent firms are active. However, even if subsequent activity is not neglected, the following result holds*:

If $A(Q_{i-1}) \neq \varnothing$, *where* $Q_{i-1} = \sum_{j=1}^{i-1} q_j$, *then firm* $i (i = 1,\dots,N)$ *can guarantee itself strictly positive profits.* (*Curtis Eaton and Robert Ware*, 1987, *prove a result reminiscent of this.*)

PROOF:

The appropriate strategy for firm i is

(1) If $P(Q_{i-1} + \alpha q^*) > P^*$, then choose $q_i = \alpha q^*$. In this case, Lemma 2 implies that

$$P(Q_N) > P^* = a\left[\frac{q_i}{\alpha}\right],$$

so that clearly $\Pi_i > 0$.

(2) If $P(Q_{i-1} + \alpha q^*) \leq P^*$ then q_i is chosen as follows. Define

$$\hat{q} = \mathrm{Max}\left\{q \mid q \in \overline{A(Q_{i-1})}\right\},$$

so that $$P(Q_{i-1}+\hat{q}) = a\left[\frac{\hat{q}}{\alpha}\right]$$

and $$\hat{q} \leq \alpha q^*,$$

where the bar denotes set closure. (Note that $A(Q_{i-1})$ is open.) Now take any

$$q_i \in A(Q_{i-1}), \text{ where } \hat{q} - q_i < \alpha\bar{q}.$$

It follows that

$$\alpha\bar{q} < q_i < \alpha q^*.$$

It now will be shown that no firm after i can be active, ensuring that $\Pi_i > 0$ since $q_i \in A(Q_{i-1})$. For suppose firm $k > i$ is active, $q_k > 0$, $\Pi_k \geq 0$. Hence

$$P(Q_{i-1}+q_k) > P(Q_{i-1}+q_i+q_k) \geq a\left[\frac{q_k}{\alpha}\right],$$

so that

$$\alpha\bar{q} < q_k < \hat{q} \leq \alpha q^*.$$

Hence

$$\begin{aligned} P(Q_{i-1}+\hat{q}) = a\left[\frac{\hat{q}}{\alpha}\right] &< a\left[\frac{q_k}{\alpha}\right] \\ &\leq P(Q_{i-1}+q_i+q_k) \\ &< P(Q_{i-1}+\hat{q}), \end{aligned}$$

a contradiction. □

A final straightforward property of any equilibrium follows, in turn from Lemma 3:

LEMMA 4: *Active Firms Are First. Consider any Stackelberg oligopoly game as in Assumptions* 1 *and* 2 *and Definitions* 2 *and* 3. *Assume that* $A(0) \neq \varnothing$. (*That is, a monopoly could make positive profits.*) *Then all active firms make positive profits and precede any and all inactive firms.*

PROOF:

Since $A(0) \neq \varnothing$, Lemma 3 implies that the first firm can guarantee itself positive profits and hence must be active.

Consider any other active firm, j, say, where $j \geq 2$. It must be the case that $A(Q_{j-1}) \neq \varnothing$ and hence $\Pi_j > 0$. For suppose that $A(Q_{j-1}) = \varnothing$ so that $\Pi_j = 0$ with $q_j > 0$. It is easily seen that firm j is then the last active firm. There must also be a last active firm *prior* to j, since firm 1 is always active. But now this firm can increase its profit by a sufficiently small increase in output because this closes firm j down. This constitutes a contradiction.

Suppose now an inactive firm, i, precedes an active firm, j. Since $A(Q_{j-1}) \neq \varnothing$ and $Q_{i-1} \leq Q_{j-1}$, it follows that $A(Q_{i-1}) \neq \varnothing$ and firm i can guarantee itself positive profits by Lemma 3, a contradiction.

This last lemma serves to place a limit on residual demand facing the last firm, given that there are enough potential firms that some must be inactive. The main result of this section can then be proved:

THEOREM 2: *Convergence to Competition. Consider any Stackelberg oligopoly game as in Assumptions* 1 *and* 2 *and Definitions* 1 *and* 2. *Consider any subgame-perfect equilibrium of this game as in Definition* 3. *Suppose that the scale parameter*, α, *tends to zero. but* $N > N(\alpha)$ *always, where Lemma* 1 *defines* $N(\alpha)$. (*Hence* $N \to \infty$.) *It follows that*

$$P(Q_N) \to P^*, \text{ where } Q_N = \sum_{j=1}^{N} q_j,$$

$$\text{so } Q_N \to Q^*.$$

Furthermore, for any given integer i, firm i is active for all small enough α, *and*

$$\frac{q_i}{\alpha} \to q^*, \text{ as } \alpha \to 0.$$

Hence the Stackelberg equilibria converge to the competitive equilibrium in the natural sense.

PROOF:

Lemmas 1 and 4 imply that, since $N > N(\alpha)$, firm N must be inactive in any subgame-perfect equilibrium as in Definitions 1, 2, and 3. Hence

$$P(Q_{N-1}) \leq P^* + \bar{s}\alpha q^*,$$

where

$$\bar{s} = \underset{Q \in [0, Q^*]}{\operatorname{Max}} |P'(Q)| > 0,$$

since if $P(Q_{N-1}) > P^* + \bar{s}\alpha q^*$, firm N could obtain positive profits by choosing

$$q_N = \alpha q^*.$$

Hence, using Lemma 2,

$$P^* < P(Q_N) \leq P(Q_{N-1}) \leq P^* + \bar{s}\alpha q^*$$

and clearly then

$$P(Q_N) \to P^* \text{ as } \alpha \to 0.$$

The number of active firms must tend to infinity as $\alpha \to 0$. (This is a stronger statement, of course, than $N(\alpha) \to \infty$, as $\alpha \to 0$). For if not, there exist active firms with average cost bounded above P^*, as $\alpha \to 0$, which is impossible given $P(Q_N) \to P^*$. Lemma 4 implies that all firms are then eventually active as $\alpha \to 0$. Finally, if it were not true that

$$\left[\frac{q_i}{\alpha}\right] \to q^* \text{ as } \alpha \to 0 \; \forall i,$$

then, again, there would exist cases where average cost exceeds $P(Q_N)$, a contradiction.

It should be noted that Marcel Boyer and Michel Moreaux (1986) independently obtain similiar results under stronger assumptions.

II. Natural Monopoly

Suppose now each firm has always decreasing average cost. Since this is claimed to be empirically sometimes plausible, this case has inherent interest. Furthermore, as discussed in the Introduction, this case is the basis of an interesting comparison among the present model, the Cournot model with free entry, and the contestability model.

Hence replace Assumption 1 with:

ASSUMPTION 3: *Cost. Each firm has identical long-run average cost given by the continuously differentiable function*

$$a: (0,\infty) \to (0,\infty),$$
$$a'(q) < 0 \quad \forall q \in (0,\infty).$$

A monotonically decreasing function which is bounded below must have a limit. Thus there must exist P^* such that

$$a(q) \to P^*, \text{ as } q \to \infty.$$

P^* is assumed strictly positive. Total cost is

$$\begin{aligned} c(q) &= 0, q = 0 \\ &= q \cdot a(q), \forall q \in (0,\infty), \end{aligned}$$

which, again, need not be continuous at $q = 0$.

Assumption 2 concerning demand is retained precisely as stated. Definition 1 describing the formal scaling-down process still captures an intuitive notion of a firm becoming smaller relative to demand. In this case, it follows that, for any fixed $q > 0$,

$$a\left[\frac{q}{\alpha}\right] \to P^*, \text{ as } \alpha \to 0.$$

Hence this process has a limit in which all firms have constant average cost at level P^*. Definitions 2 and 3 can be retained here, given that Assumption 3 above replaces Assumption 1. Lemma 1 is still valid as stated and as proved. Lemma 2 is still valid as stated, and indeed the proof is now trivial. Lemma 4 also remains valid. Lemma 3 can be strengthened as:

LEMMA 5: *Sufficient Condition for Activity. Suppose Assumptions* 3 *and* 2 *and Definitions* 1, 2, *and* 3 *describe a Stackelberg oligopoly* *game. As before, define the set, for all* $Q \geq 0$

$$A(Q) = \left\{ q | P(Q+q) > a\left[\frac{q}{a}\right] \right\}$$

Then, if $A(Q_{i-1}) \neq \varnothing$,

where $$Q_{i-1} = \sum_{j=1}^{i+1} q_j,$$

firm i can guarantee itself a strictly positive profit and simultaneously deter all further activity. (This is not a claim, as yet, that such a choice by firm i is optimal.)

PROOF:

The appropriate strategy is that defined for possibility "(2)" in the proof of Lemma 3, where all conditions involving q^* should be interpreted as automatically satisfied.

In such an natural monopoly situation, a Stackelberg equilibrium with free entry can be explicitly constructed. This equilibrium has a striking structure. There is a single active firm, as would be true with contestability, but not usually with Cournot. This also contrasts with Avinash Dixit (1979) where, with decreasing average cost, the Stackelberg leader might choose to permit a single follower to be active. That is, a larger number of potential firms can lead to a smaller number of active firms.

THEOREM 3: *Only First Firm Active. Suppose Assumptions* 3 *and* 2 *and Definitions* 1, 2, *and* 3 *describe the Stackelberg oligopoly game. Suppose* $N > N(\alpha)$, *where Lemma* 1 *defines* $N(\alpha)$, *so some firms are inactive. Suppose* $A(0) \neq \varnothing$, *so a monopolist could make a positive profit. Then, in any subgame-perfect equilibrium, only the first firm is active.*

It should be noted that Marcel Boyer and Michel Moreaux (1986) independently obtain related results under stronger assumptions.

PROOF:

Since $A(0) \neq \varnothing$, firm 1 can guarantee itself strictly positive profits and deter all further entry, by Lemma 5.

Suppose now there is an SPE with firms other than firm 1 active. Firm 1 must still obtain strictly positive profits, or the strategy above would be preferable. Hence

$$\begin{aligned} 0 &< q_1\left[P(Q_N) - a\left[\frac{q_1}{\alpha}\right]\right] \\ &< Q_N\left[P(Q_N) - a\left[\frac{q_1}{\alpha}\right]\right] \\ &< Q_N\left[P(Q_N) - a\left[\frac{Q_N}{\alpha}\right]\right], \end{aligned}$$

since $0 < q_1 < Q_N$. Thus firm 1 can obtain greater profits by producing output Q_N instead of q_1, supposing other firms are inactive. But Q_{N-1} must be such that firm N can obtain only zero profits, by Lemma 4, since $N > N(\alpha)$. Hence $(Q_N + \varepsilon)$ will strictly deter entry for any $\varepsilon > 0$. Since profit is continuous function of positive output, choosing ε small enough guarantees that $(Q_N + \varepsilon)$ is still preferable for firm 1, contradicting the assumption that other firms could be active in some SPE.

Such SPE can be constructed as follows. Strategies for firms $i = 1, \ldots, N$ are

(i) If $A(Q_{i-1}) = \varnothing$, then $q_i = 0$

(ii) If $A(Q_{i-1}) \neq \varnothing$, then choose q_i as any solution of

$$\operatorname*{Max}_{q_i} \left[P(Q_{i-1} + q_i) - a\left[\frac{q_i}{\alpha}\right]\right] q_i$$

$$\text{s.t. } A(Q_{i-1} + q_i) = \varnothing.$$

Lemma 5 implies that if $A(Q_{i-1}) \neq \varnothing$, then it is possible to make a *strictly* positive profit and *guarantee* no further activity. Note that if $A(Q_{i-1}) \neq \varnothing$, the above strategy yields at least as much as profit. Finally, by continuity, the set

$$\{Q | A(Q) = \varnothing\} = [\bar{Q}, \infty)$$

for some $\bar{Q}$. Hence the problem under (ii) has a solution and yields strictly positive profits.

It is not hard to demonstrate that these strategies do constitute an SPE, and that furthermore there are no other SPEs. Note that these strategies entail firms shutting down in the case that activity yields zero maximum profits. □

The above Stackelberg equilibrium with free entry will usually be a little less efficient than the outcome predicted by the contestability model. This follows because this latter model involves a single active firm producing maximum output subject to nonnegative profits. However, it would seem that this Stackelberg equilibrium with one active firm is likely to be more efficient than a Cournot equilibrium with, usually, more than one, given decreasing average cost. A factor complicating this comparison is that the first Stackelberg firm will usually set a quantity barely sufficient to deter other activity, whereas a Cournot equilibrium with free entry will usually involve a total output level greater than this. To see how these effects might interact, a simple example is considered. This example is simple enough that a Cournot equilibrium always exists, although free entry does not always determine the number of active firms uniquely.

LEMMA 6: *Stackelberg Versus Cournot Efficiency. Suppose average cost for each firm is*

$$a(q) = \frac{F}{q} + m; \quad F, m > 0,$$

so that Assumption 1 *is satisfied, with* $P^* = m$. *Demand is given as*

$$\begin{aligned} P &= a - Q, Q \le a \\ &= 0, \qquad Q > a, \end{aligned}$$

so that Assumption 2 *is satisfied, given* $a > m$. *Then*:

(*i*) *If a Cournot equilibrium has one firm, this is also the Stackelberg equilibrium.*

(*ii*) *If a Cournot equilibrium has two firms, it may or may not be more efficient than the Stackelberg equilibrium, in a total surplus sense.*

(*iii*) *If a Cournot equilibrium has three or more firms, it is necessarily less efficient than the Stackelberg equilibrium, in this total surplus sense.*

PROOF:

See the Appendix. □

Presumably the Cournot model and the contestability model will predict asymptotic efficiency when the scale of each firm is shrunk relative to demand. This is certainly true for the present model, where *potential* competition ultimately drives even the single active firm to the efficient level of output:

THEOREM 4: *Convergence to Efficient Output. Suppose that Assumptions* 3 *and* 2 *and Definitions* 1, 2, *and* 3 *describe the Stackelberg oligopoly game. Then, as* $\alpha \to 0$, *with* $N > N(\alpha)$ *always, it follows that*

$$P(Q_N) \to P^*, Q_N \to Q^*,$$

where, in light of Theorem 3

$$Q_N = Q_1 = q_1$$

is the output produced by the first firm.

PROOF:

It will be shown that, if $\bar{P} > P^*$, there exists $\bar{\alpha} > 0$ such that if $\alpha \in (0, \bar{\alpha}]$, and $N \geq N(\alpha)$, then any SPE yields $P(Q_N) \in (P^*, \bar{P}]$.

It is clear here that $P(Q_N) > P^*$, otherwise all firms make losses. Suppose, however, it is possible to find arbitrarily small α such that, although $N > N(\alpha)$, there exists some SPE with

$$P(Q_N) > \bar{P}.$$

Consider the situation facing firm 2. Define

$$\hat{q} = \frac{\bar{P} - P^*}{2\bar{s}},$$

where, as before,

$$\bar{s} = \max_{Q \in [0, Q^*]} |P'(Q)|.$$

It follows that

$$\hat{P} = P(q_1 + \hat{q}) \geq \tfrac{1}{2}(\bar{P} + P^*) > P^*.$$

From Assumption 3 it follows that there is an $\hat{\alpha} > 0$ such that

$$0 < \alpha < \hat{\alpha} \text{ implies } a\left[\frac{\hat{q}}{\alpha}\right] < \hat{P},$$

which contradicts Theorem 3 and Lemma 5, since now $A(q_1) \neq \varnothing$.

III. Empirical Evaluation

The present paper develops the properties of a generalized Stackelberg model of oligopoly. An attempt is made to argue that this model compares favorably both with the Cournot model and with the contestability model. This is done by considering the plausibility of the assumptions and results of the three models. Of course, empirical evaluation is ultimately the most meaningful test. The Stackelberg model might be favored to the extent that actual industries can have asymmetric output patterns despite similar costs for the firms.

Finally, it should be admitted that none of the three models here does justice to the multitude of dynamic interactions which could arise in an oligopoly, at least in theory. However, the game-theoretic models which do account more fully for these possess riches sufficient to embarrass King Croesus. Not only do such models, that is, predict almost anything can be an equilibrium outcome but most also do not address the issue of convergence to competition. (Ariel Rubinstein, 1979, for example, provides an elegant proof of the "perfect folk theorem." It is clearly not the intention in this context to discuss such convergence to competition, but this theorem could be interpreted as allowing collusion in an arbitrarily large cartel.) Such game theory models, to date at least, are then scarcely "falsifiable" to use Karl Popper's (1965) term.

APPENDIX

Example A

This was mentioned in the Introduction. The point is to provide a neutral setting for a "survival contest" between Cournot equilibria and Stackelberg equilibria. The Stackelberg equilibria may emerge as "fittest," in the sense of corresponding to the only possible pure strategy equilibria.

Two firms produce an homogeneous good whose inverse demand is

$$P = a - q_1 - q_2,$$

where P is price and $q_i, i = 1,2$ are the respective quantities. This market exists only at $t = 0$, but there is also a prior period, $t = -1$. Production uses only capital and is constant returns to scale. Each firm can purchase capital at $t = -1$ or at $t = 0$ or both, but cannot sell capital once acquired. The amount of capital firm i buys at $t = -1$ is $\bar{k}_i$ and at $t = 0$ is $\tilde{k}_i$. Output at $t = 0$, q_i, cannot exceed total capital. (Normalizing units). The price of capital is $c, c < a$, and the interest rate is $r > 0$, so each firm has profits

$$\Pi_i = (a - q_1 - q_2)q_i - c(1+r)\bar{k}_i - c\tilde{k}_i,$$

where

$$0 \leq q_i \leq \bar{k}_i + \tilde{k}_i; \quad \bar{k}_i, \tilde{k}_i \geq 0 \qquad i = 1,2.$$

Subgame perfectness requires that there be a Nash equilibrium at $t = 0$ in $(\tilde{k}_1, q_1)$ and $(\tilde{k}_2, q_2)$ for *any* given values of $\bar{k}_1$ and $\bar{k}_2$. Note first that if $\tilde{k}_i > 0$, so *some* capital is acquired at $t = 0$, then

$$q_i = \bar{k}_i + \tilde{k}_i,$$

for otherwise firm i could gain by reducing $\tilde{k}_i$. In this case, then, there can be no "excess capacity." Hence $\tilde{k}_i$ can be eliminated from the expression for Π_i to yield

$$\Pi_i = \{(a - q_1 - q_2)q_i - c\,\mathrm{Max}[q_i - \bar{k}_i, 0]\}$$
$$- c(1+r)\cdot\bar{k}_i,$$

where the last term is fixed at $t = 0$. Consider then firm i's best reply function for choice of q_i at $t = 0$. This is:

$$q_i = \begin{cases} 0, & \text{if } q_j \geq a \\ \frac{1}{2}(a - q_j), & \text{if } a \geq q_j \geq a - 2\bar{k}_i \end{cases} \text{“}A\text{”}$$
$$\qquad \begin{cases} \bar{k}_i, & \text{if } a - 2\bar{k}_i \geq q_j \\ & \quad \geq a - c - 2\bar{k}_i \\ \frac{1}{2}(a - c - q_j), & \text{if } a - c - 2\bar{k}_i \geq q_j \quad \text{“}B\text{”} \end{cases}$$

Firm j has an analogous best-reply function and it is easy to show graphically that a unique Nash equilibrium in $(\tilde{k}_1, q_1)$ and $(\tilde{k}_2, q_2)$ exists at $t = 0$ for all $\bar{k}_1$ and $\bar{k}_2$.

Now consider the situation facing the firms at $t = -1$. Each firm must select $\bar{k}_i$ bearing in mind the effect this choice has on the equilibrium at $t = 0$. Indeed, $\bar{k}_1$ and $\bar{k}_2$ must, themselves, be equilibrium choices. Suppose that, in equilibrium,

$$0 \le q_i < \bar{k}_i,$$

so there is excess capacity at $t = 0$. The equilibrium at $t = 0$ must lie in region "A" of firm i's best-reply function above. But then reducing $\bar{k}_i$ reduces firm i's cost but does not affect the equilibrium at $t = 0$ or firm i's revenue.

Hence

$$q_i \ge \bar{k}_i \quad i = 1,2,$$

and there is no excess capacity at all in equilibrium. (Such a point was previously made by Avinash Dixit, 1980). Suppose now in equilibrium

$$q_i > \bar{k}_i > 0,$$

so capital is purchased in both periods. This means the equilibrium at $t = 0$ must lie somewhere in region "B" of firm i's best-reply function above. Again reducing $\bar{k}_i$ does not affect the $t = 0$ equilibrium or firm i's revenue. However, it reduces firm i's cost because purchase of some capital is shifted forward in time. Hence

$$q_i > \bar{k}_i \text{ implies } \bar{k}_i = 0 \quad i = 1,2$$

and firm i either purchases all capital at $t = -1$ or all of it at $t = 0$.

The foregoing observations limit the pure-strategy equilibria to the three types below:

(i) Both $\bar{k}_i > 0$, with then $q_i = \bar{k}_i$ also. In this case, either firm can increase profits by setting $\bar{k}_i = 0$ but purchasing some $\bar{k}_i > 0$. Hence simultaneous choice of capital at $t = -1$ is never an equilibrium, given $r > 0$.

(ii) Both $\bar{k}_i = 0$, so that

$$\Pi_i = (a - q_1 - q_2)q_i - cq_i.$$

Clearly this yields equilibrium outputs and profits

$$q_i = \left(\frac{a-c}{3}\right), \Pi_i^c = \left(\frac{a-c}{3}\right)^2, i = 1,2.$$

(iii) Firm i chooses $q_i = \bar{k}_i > 0$, but j chooses $\bar{k}_j = 0$. Firm j must choose in accordance with

$$q_j = \begin{cases} 0, & \text{if } \bar{k}_i > (a-c), \\ \dfrac{a-c-\bar{k}_i}{2} & \text{if } \bar{k}_i \le (a-c). \end{cases}$$

Substituting this into Π_i and maximizing over $\bar{k}_i$ yields

$$\bar{k}_i = \frac{a-c-2cr}{2}, \quad \text{if} \quad 0 < r < \frac{a-c}{2c}$$

and

$$\Pi_i^L = \frac{(a-c-2cr)^2}{8}.$$

In this equilibrium, it will never pay firm j to select $\bar{k}_j > 0$.

(ii) versus (iii). There are two equilibria of type (iii) and none of type (ii) if

$$\Pi_i^L > \Pi_i^c \quad i = 1,2,$$

which clearly holds when r is small enough. Indeed, as $r \to 0$,

$$q_i = \bar{k}_i \to \frac{a-c}{2}, \qquad q_j \to \frac{a-c}{4},$$

the usual Stackelberg choices.

Example B

A computer program was developed to solve numerically for a subgame perfect equilibrium of such Stackelberg oligopoly games by means of "backward recursion." That is, the first step was to derive the last firm's best-reply to any possible total quantity produced by the previous firms. Given this, the second-to-last firm's best-reply could then be constructed, and so on. Each firm here was limited to a quantity choice which was a multiple of some grid size parameter.

For example, suppose average cost is

$$a(q) = \frac{1}{2q} + \frac{q}{2} \quad \forall q > 0,$$

which satisfies Assumption 1, with $P^* = q^* = 1$. Suppose inverse demand is

$$P(Q) = 2 - Q,$$

which then satisfies Assumption 2, with $Q^* = 1$. Suppose each firm is allowed to choose multiples of 1/2000 in [0,1]. When $\alpha = \frac{1}{4}$, the following results were obtained for the free-entry case

$$P(Q_N) = 1.224, \; Q_N = 0.776$$

$$q_1 = 0.2800, \; q_2 = 0.2515, \; q_3 = 0.2445$$

$$\pi_1 = 0.06092, \; \pi_2 = 0.05633, \; \pi_3 = 0.05471$$

with all subsequent firms inactive. In this case, then, both quantity and profit strictly decrease with the index of the firm. Indeed, firm 3 produces less than $\alpha q^* = 0.25$. When $\alpha = \frac{1}{8}$, the following results were obtained with free entry

$$P(Q_N) = 1.118, \quad Q_N = 1.082$$

$$q_1 = q_2 = q_3 = q_4 = q_5 = 0.1400, \quad q_6 = 0.1820$$

$$\pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = 0.01562, \quad \pi_6 = 0.008480$$

with all subsequent firms inactive. It is interesting to observe that the last active firm, firm 6, produces more than any of the previous active firms, although it earns less profit. It seems that the explanation is that this firm is the first firm to find it worthwhile to keep all subsequent firms inactive, by means of a relatively high output. Furthermore, small changes in the output produced by any previous firm would then be exactly offset by this last firm. Each of these previous firms would then face a locally fixed price. Indeed, it is readily checked that for firms 1,...,5

$$P = 1.118 \simeq \mathrm{MC} = 8 \times 0.140 = 1.120.$$

This suggests that the Stackelberg equilibrium with free entry might converge relatively rapidly to an efficient allocation, although theoretical work is needed to prove this.

It would also be of interest to try to obtain an example in which profit, and not merely quantity, *increases* with the index of the firm.

This computer program was written with an eye on the empirical testing of the Stackelberg model, as described in Section III.

PROOF OF LEMMA 6:

(i) If a Cournot equilibrium has one firm, it must produce monopoly output, $\frac{1}{2}(a-m)$, and this must suffice to deter entry. But then the Stackelberg equilibrium's first firm will also choose this quantity.

(ii), (iii) If n is the number of active firms in a Cournot equilibrium, then output per firm is

$$q_i = \frac{a-m}{n+1}$$

so that total output and price are

$$Q = \frac{n(a-m)}{n+1}, \quad P = \frac{a}{n+1} + \frac{nm}{n+1}.$$

Since active firms earn nonnegative profits, but inactive firms must be choosing optimally, it follows that

$$\frac{(a+nm)(a-m)}{4(n+1)^2} \le \alpha F \le \frac{(a+nm)(a-m)}{(n+1)^2}.$$

The Cournot surplus is

$$\begin{aligned} TCS &= (a-m)Q - \tfrac{1}{2}Q^2 - n\alpha F \\ &= \frac{(n^2+2n)(a-m)^2}{2(n+1)^2} - n\alpha F. \end{aligned}$$

Consider now the Stackelberg equilibrium from Theorem 3. Here

$$\begin{aligned} q_1 = Q_1 = Q_N &= \mathrm{Max}\left\{\frac{a-m}{2}, a-m-2\sqrt{\alpha F}\right\} \\ &\ge a-m-2\sqrt{\alpha F}. \end{aligned}$$

(This expression shows that, as $\alpha \to 0$, $Q_N \to a-m$, the efficient level, as predicted by Theorem 4.) Now the total Stackelberg surplus is

$$\begin{aligned} TSS &\ge (a-m)(a-m-2\sqrt{\alpha F}) \\ &\quad - \tfrac{1}{2}(a-m-\sqrt{\alpha F})^2 - \alpha F \\ &= \frac{(a-m)^2}{2} - 3\alpha F \end{aligned}$$

Hence it follows that

$$TSS - TCS \ge \frac{(a-m)^2}{2(n+1)^2} + (n-3)\alpha F > 0 \quad \text{if } n \ge 3$$

establishing (iii).

Suppose then $n = 2$, as in (ii). For this to be a Cournot equilibrium

$$\frac{(a+2m)(a-m)}{36} \le \alpha F \le \frac{(a+2m)(a-m)}{9}.$$

To demonstrate that both possibilities exist, it is enough to consider the case that monopoly output does not deter entry, so that

$$\alpha F \le \frac{(a-m)^2}{16}.$$

In this case, the lower bound on TSS holds with equality and

$$TSS - TCS = \frac{(a-m)^2}{18} - \alpha F,$$

which is easily seen to be ambiguous in sign even in the light of the upper and lower bounds on αF obtained above. (Consider the case where m is small, for example.) Hence (ii) is proved.

REFERENCES

Baumol, William J., "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review*, March 1982, *72*, 1–15.

———, **Panzar, John C. and Willig, Robert D.**, *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.

———, "Contestable Markets: An Uprising in the Theory of Industry Structure: Reply," *American Economic Review*, June 1983, *73*, 491–96.

Boyer, Marcel and Moreaux, Michel, "Perfect Competition as the Limit of a Hierarchical Market Game," *Economic Letters*, 1986, *22*, 115–18.

Brock, William A. and Scheinkman, José A., "Free Entry and the Sustainability of Natural Monopoly: Bertrand Revisited by Cournot" in E. Evans, ed., *Breaking Up Bell: Essays on Industrial Organization and Regulation*, New York: North-Holland, 1983.

Dixit, Avinash, "A Theory of Duopoly Suggesting a Theory of Entry Barriers," *Bell Journal of Economics*, Spring 1979, *10*, 20–32.

______, "The Role of Investment in Entry Deterrence," *Economic Journal*, March 1980, *90*, 95–106.

Eaton, B. Curtis and Ware, Robert, "A Theory of Market Structure with Sequential Entry," *Rand Journal of Economics*, Spring 1987, *18*, 1–16.

Edgeworth, F. Y., "The Pure Theory of Monopoly," in *Papers Relating to Political Economy*, Macmillan: London, 1925, 111–42.

Fudenberg, Drew and Tirole, Jean, "Preemption and Rent Equalization in the Adoption of New Technology," *Review of Economic Studies*, July 1985, *52*, 383–402.

Hellwig, Martin, Leininger, Wolfgang, Reny, Philip J. and Robson, Arthur J., "Subgame Perfect Equilibrium in Continuous Games of Perfect Information: An Elementary Approach to Existence and Approximation by Discrete Games," *Journal of Economic Theory*, forthcoming 1990.

McManus, Maurice, "Equilibrium, Number and Size in Cournot Oligopoly," *Yorkshire Bulletin of Social and Economic Research*, November 1964, *16*, 68–75.

Novshek, William, "Cournot Equilibrium with Free Entry," *Review of Economic Studies*, April 1980, *47*, 473–86.

Popper, Karl R., *The Logic of Scientific Discovery*, New York: Harper & Row, 1965.

Roberts, John and Sonnenschein, Hugo, "On the Foundations of the Theory of Monopolistic Competition," *Econometrica*, January 1977, *45*, 101–113.

Rubinstein, Ariel, "Equilibrium in Supergames with the Overtaking Criterion,' *Journal of Economic Theory*, June 1979, *21*, 1–9.

Schwartz, Marius and Reynolds, Robert J., "Contestable Markets: An Uprising in the Theory of Industry Structure: Comment," *American Economic Review*, June 1983, *73*, 488–90.

Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, 25–55.

Spence, Michael A., "Entry, Capacity Investment and Oligopolistic Pricing," *Bell Journal of Economics*, Autumn 1977, *10*, 1–19.

Stackelberg, H. von, "Marktform und Gleichgewicht," Berlin and Vienna: Springer, 1934.

Weitzman, Martin, L., "Contestable Markets: An Uprising in the Theory of Industry Structure: Comment," *American Economic Review*, June 1983, *73*, 486–87.

Quality vs. Quantity in Military Procurement

By WILLIAM P. ROGERSON*

It is often argued that the same expenditures on military procurement would produce a more effective defense if larger numbers of less sophisticated (and thus cheaper) weapons were purchased. This paper shows that such a result can occur even if the military derives no private consumption value from technically sophisticated weapons. Rather the organization of the decision-making process itself can produce this result. This suggests some possible solutions through organizing decision making in a different fashion. (JEL 022)

Many institutional analyses of defense procurement raise the issue that the military's choice along the quality-quantity frontier for various weapons systems seems to be biased toward "too high" a level of quality. That is, it is argued that the same expenditures would produce a more effective defense if larger numbers of less elaborate and less technically sophisticated (and therefore cheaper) weapons were purchased. Frederic Scherer (1964, pp. 29–38), for example, documents the military's apparent bias toward quality in a long series of examples. He concludes that "the competitive incentives in weapons acquisition in the 1950s ran strongly in favor of quality maximization and lead time minimization, but almost negligibly in favor of development cost minimization."[1] Similar points are made by Jacques Gansler (1980, pp. 15–21), Merton Peck and Frederic Scherer (1962, ch. 13), Richard Stubbing (1986, ch. 8) and Sheila Tobias et al. (1982, pp. 366–78).

Critics of the current approach cite the evidence provided in 1978 by a series of mock air battles organized by the air force and navy called *ACEVAL/AIMVAL*. In these battles, pilots flying the F5E were pitted against pilots flying the F14 and F15. The F5E is a small inexpensive fighter designed primarily for export to other countries. The unit cost is approximately $5 million. The F14 and F15 are the navy and air force's premiere fighters. Their unit costs are, respectively, approximately $47 million and $30 million.[2] Thus the premiere fighters cost 6 to 9 times as much as the simple fighter. Tobias et al (1982, p. 216) sum up the results of the test as follows:

> "In one-on-one encounters between planes equipped with heat-seeking missiles such as the sidewinder, planes tended to destroy each other (so lethal are these weapons). In two-against-one encounters, no matter how sophisticated the single plane, the two always defeated the one. These results have shaken some of previously held assumptions about the advantages of very expensive, very sophisticated high-performance aircraft."

The standard explanation for this bias toward quality has been that military decision makers and design engineers derive private "consumption value" from creating and fielding technologically advanced systems.[3] Thus the problem is seen to be that military decision makers reduce military preparedness in order to indulge their own private interests in creating new technological ad-

---

*Department of Economics, Northwestern University, 2003 Sheridan Road, Evanston, IL 60208. I would like to thank Tai Yeong Chung, Kathleen Hagerty, and an anonymous referee for extremely helpful discussions and comments. This work was supported by NSF grant no. SES-8705477 and the Rand Corporation through funds provided by Program Analysis and Evaluation, Office of the Secretary of Defense.

[1] Scherer (1964), p. 36.

[2] These figures are from Tobias et al. (1982), pp. 211, 213 and 217.

[3] See, for example, Gansler (1980) pp. 101 and 103.

vances. There is undoubtedly some truth to this claim.

The purpose of this paper, however, is to advance a different theory explaining this bias which does not rely on the existence of private "consumption values" of military decision makers. It will be shown that the institutional organization of decision making may in and of itself result in excessive quality even when the military derives no private value from quality. Thus even agents operating in good faith to maximize military preparedness may choose too high a quality and too low a quantity depending on how the decision-making hierarchy is organized.

The standard theory sees the problem as originating from "bad" agents and suggests few solutions other than attempting to get "better" agents or monitoring the "bad" ones more closely. In contrast, this paper's theory sees the problem as originating from the manner in which decision making is organized. This suggests that solutions may be obtainable by reorganizing the decision-making process. In particular the theory suggests two types of changes which may help ameliorate the biases generated by the current organization. These are precommitment to fixed budget levels and use of overlapping jurisdictions between military services.

The model can also be interpreted as applying to any organization engaged in a complex production task where design or project selection decisions must be delegated to subordinates. Managers may choose overly high quality levels for their projects even though they derive no consumption value from quality and are acting in good faith to increase the effectiveness of their programs.

I am not aware of any formal theoretical work which is closely related to this paper. The typical paper found in the procurement and regulation literature treats the procurer (Congress) as a Stackleberg leader who presents the agent (the military) with a schedule of options to choose from. In this paper, Congress has neither the information nor the commitment ability to act as a Stackleberg leader. Instead, Congress simply reacts to military proposals for new weapons by choosing the level of funding which maximizes its perceived benefits from acquiring the weapon. Thus the standard methodology of a one-dimensional self-selection model plays no role in this paper. A second difference of this paper is that it focuses on the relationship between Congress and the military. The bulk of formal models of procurement analyze the relationship between the military and its contractors.[4]

Section I describes the model and key assumptions which generate the results. Section II analyzes the model and Sections III and IV explore two possible solutions to the problem. Section V argues that the model can be interpreted as applying to any complex organization and Section VI provides a brief conclusion.

I. The Model

A. *The Optimal Program*

A weapons program will be assumed to be completely described by two nonnegative real numbers, (q, x), where q denotes the quality of the weapon and x denotes the number of units purchased. Let $C(q, x)$ denote the dollar cost of producing a weapons system of quality q in amount x. Let $V(q, x)$ denote the military value of the weapons program (q, x) to Congress, expressed in dollars. The optimal weapons program from Congress's point of view[5] therefore maximizes benefits minus costs. This will also be called the first-best program.

Definition 1: The program (q, x) is optimal or first-best if it satisfies

$$\text{(1)} \quad \underset{(\hat{q},\hat{x}) \in R^{+2}}{\text{Maximize}} \; V(\hat{q},\hat{x}) - C(\hat{q},\hat{x}).$$

---

[4] See David Baron and David Besanko (1987), Preston McAfee and John McMillan (1986), Michael Riordan and David Sappington (1986), and Jean Tirole (1986), for examples of the standard approach. Also see Tracy Lewis (1986) for a different model of the relationship between Congress and the military.

[5] Congress's preferences will be interpreted as society's preferences. Therefore the optimal program from Congress's point of view will be spoken of simply as the optimal program.

It will also be useful to define a second-best program for any given budget level.

Definition 2: The program (q, x) is a second-best program given the budget B if it satisfies

$$\text{(2)} \qquad \underset{(\hat{q},\hat{x}) \in R^{+2}}{\text{Maximize}}\, V(\hat{q},\hat{x})$$

$$\text{(3)} \qquad \text{subject to } C(\hat{q},\hat{x}) \leq B.$$

B. *The Equilibrium Program*

The decision-making process leading up to the adoption of a new weapons program will be viewed as a two-person game between Congress and the military. Two fundamental assumptions about the nature of this game will be made. These will now be described.

The first fundamental assumption concerns the nature of the military's preferences. Let $>_m$ denote the military's preference ordering over weapons programs. It will be assumed that the military's preferences are represented by $V(q, x)$. That is,

$$\text{(4)} \quad (q,x) >_m (\hat{q},\hat{x}) \Leftrightarrow V(q,x) > V(\hat{q},\hat{x}).$$

Note that there is a sense in which the military is acting as a "good faith" agent for Congress. It cares solely about the military value of programs and it agrees with Congress on the military value of various possible programs. In particular the military derives no private consumption value from quality.

However, the military is not a perfect agent. A perfect agent's preferences would be defined by

$$\text{(5)} \quad (q,x) >_m (\hat{q},\hat{x}) \Leftrightarrow V(q,x) - C(q,x) > V(\hat{q},\hat{x}) - C(\hat{q},\hat{x}).$$

Presumably the cost of a program, $C(q, x)$, is fairly easy to objectively measure. If it was equally easy to objectively measure the dollar social value of a program, $V(q, x)$, then Congress could probably induce the military to have preferences given by (5). Rewards could be made to vary directly with the size of $V - C$. However, the dollar social value of a weapons program is clearly not objectively measurable. Given this it may be difficult for Congress to directly induce preferences through incentive contracts which are perfect in the sense of (5). Military decision makers' own personal preferences may therefore play a role in shaping the nature of $>_m$.

This paper assumes that the military is as idealistic as one could reasonably expect. The reason for this is that the point of the paper is to show that the inherent structure of the decision-making process will generate a decision-bias in the quality-quantity input mix even when the military is as idealistic as possible.

In what sense are the preferences defined by (4) as idealistic as realistically possible? One might reasonably expect even a fairly idealistic agent to derive more satisfaction from increasing the contribution his project makes than from reducing its cost. An idealistic military might therefore be expected to want to maximize military preparedness. In particular, an idealistic military would not purposely reduce military preparedness solely for the sake of pursuing technology or solely for the sake of increasing its budget. However one might presumably expect an idealistic military to be in favor of increasing military preparedness even if on balance the dollar social benefits were less than the dollar social cost. Thus even an idealistic military is likely to weight maximization of $V(q, x)$ more highly than minimization of $C(q, x)$. In order to simplify the exposition, this paper makes the formal assumption that the military places no weight at all on cost minimization. However it is straightforward to show that the analysis and conclusions of this paper apply to the more general case where the military's utility function is

$$\text{(6)} \qquad V(q,x) - \lambda C(q,x),$$

where $0 \leq \lambda < 1$.

The second fundamental assumption made by this paper regards the sequential nature of the decision process. It is assumed that the military first chooses the quality of the program. Then Congress chooses the output level, given the predetermined quality. This assumption reflects two features of the procurement process. The first feature is that the technological characteristics of a weapon are determined in a development phase which precedes the production phase. The second feature is that only the military has the expertise to fully evaluate, choose, and determine the technological characteristics of the weapon. Thus Congress must delegate this decision to the military. In the stylized model of this paper, "quality" is the variable describing the technological characteristics of the system. Thus the development phase consists of the military choosing a quality. In the production phase Congress is faced with the weapon system as designed by the military. By choosing a funding level for production, Congress determines how many units will be produced.

Another way to describe the second fundamental assumption of this paper is that Congress is incapable of evaluating how $V-C$ would vary for different values of q. Congress can only evaluate how $V-C$ varies as x varies for the value of q described to it by the military. Since Congress cannot independently conceive of weapons systems of varying designs it must simply rely on the military to present a design and then decide how many to purchase.

Note that in reality Congress probably has some influence on the design choice as well as the output choice since information on design choice is not totally in the hands of the military. This paper makes the extreme assumption that the military has total control over the quality choice in order to most clearly illustrate the effects of this asymmetry of information and expertise.

The two fundamental assumptions described above result in the following structure to the game between Congress and the military. As usual, it is most convenient to think about this sort of game by beginning at the end and working backward. At stage 2 Congress chooses an output to maximize $V(q,x)-C(q,x)$ taking the military's choice of q as given. Then at stage 1 the military chooses q to maximize $V(q,x)$, realizing how the choice of q will affect Congress's choice of x.

Formally, then, an equilibrium outcome is defined as follows.

Definition 3: An equilibrium program satisfies the following:

$$(7) \qquad \underset{(q,x)\in R^{+2}}{\text{Maximize}}\, V(q,x)$$

(8) subject to

$$x \in \underset{\hat{x}\geq 0.}{\text{argmax}}\, V(q,\hat{x})-C(q,\hat{x})$$

C. *Technical Assumptions*

The various optimization problems described above will be assumed to be "well-behaved." Specifically, the following four assumptions will be made. They will be discussed after their formal statement.

ASSUMPTION 1: *V and C are twice continuously differentiable over* $(0,\infty)^2$. *All first derivatives are nonnegative. The* iso-*utility curves for V are strictly decreasing and strictly convex. The* iso-*cost curves for C are strictly decreasing and strictly concave. Finally, the function* $V-C$ *is strictly concave over* $(0,\infty)^2$.

ASSUMPTION 2: *A unique optimal program exists. Let* (q^*,x^*) *denote this program.*

ASSUMPTION 3: *For every* $q\geq 0$ *there is a unique solution to* (8). *Let* $\phi(q)$ *denote the solution to* (8) *given* q. *Assume that* $\phi(q)$ *is continuously differentiable.*

ASSUMPTION 4: *A unique equilibrium program exists. Let* (q^e,x^e) *denote this program.*

Assumptions 1 and 2 are straightforward. An immediate consequence of Assumption 1 is that unique second-best programs exist for every budget level. Let $(q^s(B),x^s(B))$ denote the unique second-best program given budget B. Assumption 3 requires that Congress always has a unique optimal output choice for any quality level chosen by the military. Assumption 4 is straightforward.

One other piece of notation will be useful. Let $E(q)$ denote the amount of money Congress will spend on the weapons program if the military chooses quality q. This is defined by

$$E(q) = C(q, \phi(q)). \tag{9}$$

II. Analysis

The purpose of this section is to investigate whether the equilibrium quality level is too large or too small. There are two possible quality levels to compare q^e with. The most obvious is the optimal quality, q^*. The second possible comparison point is the quality level of the second-best program given the equilibrium budget level. Formally, this comparison point is defined by

$$q^s(C(q^e, x^e)). \tag{10}$$

Under the second comparison, the equilibrium quality is too large (too small) if a greater level of military preparedness could be achieved with the same budget by choosing less (more) quality and more (less) output.

Clearly both comparisons are interesting. Fortunately there is no need to conduct two separate analyses. This paper's problem is well-behaved enough that both comparisons yield the same answer. That is, the equilibrium quality is too large (too small) relative to the first-best quality if and only if it is too large (too small) relative to the second-best quality. This is proven in Proposition 1. Since the proof is straightforward but somewhat long it will not be included.[6]

PROPOSITION 1:

$$q^e \gtreqless q^* \Leftrightarrow q^e \gtreqless q^s(C(q^e, x^e)). \tag{11}$$

Because of Proposition 1 the following analysis will restrict itself to comparing equilibrium quality with the first-best quality. By Proposition 1 the results also apply to the second-best comparison.

It is clear from the structure of the military's optimization problem that the equilibrium quality will not generally be first-best. The military's objective function is

$$M(q) = V(q, \phi(q)). \tag{12}$$

Thus the equilibrium quality satisfies

$$M'(q^e) = 0. \tag{13}$$

The optimal quality maximizes the following objective function:

$$M(q) - E(q). \tag{14}$$

Thus the optimal quality satisfies

$$M'(q^*) - E'(q^*) = 0. \tag{15}$$

Therefore the optimal quality, q^*, will be an equilibrium only if $E'(q^*) = 0$ which would occur only by coincidence. If $E'(q^*)$ is positive, then M is locally increasing in quality and the military will locally desire to increase quality above the first best. If $E'(q^*)$ is negative, then M is locally decreasing in quality and the military will locally desire to decrease quality below the first best. If M is single-peaked the local results are of course global. This will be stated as Proposition 2 below.

PROPOSITION 2: *Suppose M is single-peaked. That is, M is strictly increasing (decreasing) for $q < (>) q^e$. Then*

$$q^e \gtreqless q^* \Leftrightarrow E'(q^*) \gtreqless 0. \tag{16}$$

PROOF:

As above. □

There is a very natural intuition for this result. Differentiate $M(q)$ to yield

$$M'(q) = V_q(q, \phi(q)) + V_x(q, \phi(q))\phi'(q). \tag{17}$$

---

[6] See Rogerson (1988) for the proof.

The effect on M of an increase in q can be broken into two parts. The first part is the direct effect. That is, an increase in q directly increases V because V_q is positive. This is the first term in (17). The second effect is the indirect effect. An increase in q causes Congress to change its choice of x which in turn affects V. This is the second term of (17). The indirect effect may be positive or negative depending upon whether an increase in q induces an increase or decrease in x.

Suppose the military has provisionally chosen q^* and is evaluating whether to increase q slightly. The direct effect of an increase in q is always positive. If increasing q would induce Congress to increase x then the indirect effect is also positive and the military will want to increase q. If an increase in q causes only a very small drop in Congress's choice of x the indirect effect will be negative but very small and the net effect of increasing q will still be positive. However, if an increase in q causes a very large drop in x, the indirect effect will be negative and large and could overwhelm the positive direct effect. In this case the military would want to reduce quality below q^* in order to induce a more-than-compensating increase in x.

Thus the military will want to increase q above q^* if x will not drop "too much" in response. Proposition 2 shows that the correct measure of "too much" is whether expenditures fall when quality increases. When quality is increased, expenditures will also increase if output stays constant or increases. Even if output drops a small amount in response to a quality increase, expenditures will still increase. However, if output drops enough in response to a quality increase then expenditures will in fact decrease.

Proposition 2 is very useful for providing an intuition explaining why the equilibrium and optimum will in general diverge. However the proposition is not cast in terms of easily observed variables since we are presumably observing equilibrium quality choices and not optimal quality choices in existing programs. Thus if we were to guess at the sign of $E'(q)$ by looking at existing programs we would be guessing the sign of $E'(q^e)$ and not $E'(q^*)$. Proposition 3 shows that information on $E'(q^e)$ can also be used to infer whether the quality choice is too large or not. In particular, Proposition 2 remains true if q^e is substituted for q^*.

PROPOSITION 3:

$$q^e \gtreqless q^* \Leftrightarrow E'(q^e) \gtreqless 0. \tag{18}$$

PROOF:

Because $V(q,x)-C(q,x)$ is globally concave in q and x, it is straightforward to show that $M(q)-E(q)$ is also globally concave in q. Therefore it is sufficient to prove that

$$M'(q^e)-E'(q^e) \lesseqgtr 0 \Leftrightarrow E'(q^e) \gtreqless 0. \tag{19}$$

This is true because $M'(q^e)=0$. □

Thus if we could estimate or guess the sign of $E'(q)$ for current programs, we could use this information to infer whether quality was too high or too low.

One factor affecting the sign of $E'(q)$ will be the substitutability of quality for quantity. If they are very good substitutes Congress will respond to an increase in quality by a large decrease in quantity (since the increased quality is a good substitute for quantity). However, if they are very poor substitutes than an increase in quality will not induce a large decrease in quantity. Therefore equilibrium quality will tend to be too large when quality and quantity are poor substitutes. In this case the military can raise quality without fearing a large drop in Congress's choice of quantity.

Thus if it is true that quality and quantity are poor enough substitutes in most weapons programs so that increases in quality do not generate large decreases in quantity, then this paper provides a theory explaining why quality is chosen to be too high even when the military derives no private consumption value from quality.

III. Fixed-Budget Levels

Suppose Congress could precommit to a fixed budget level of B for a program prior to the point in time when the military chooses quality. That is, regardless of the quality chosen by the military, Congress commits to spend B dollars on the program. In this case the military would choose q and x to solve the following problem[7]

$$\text{(20)} \qquad \underset{q,x}{\text{Maximize}}\ V(q,x),$$

$$\text{(21)} \qquad \text{subject to } C(q,x) \leq B.$$

The solution is of course the second-best program given the budget B. Therefore, in particular, if Congress chose a budget level equal to the cost of the first-best program, the military would choose the first-best program. Let B^* denote the first-best budget level. It is defined by

$$\text{(22)} \qquad B^* = C(q^*, x^*).$$

The results of the above paragraph are summarized in Proposition 4.

PROPOSITION 4:

(1) *If Congress precommits to a budget B, the military will choose the second-best program given B.*

(2) *Therefore if Congress precommits to B^*, the military will choose the first-best program.*

PROOF:

As above. □

It is tempting to view Proposition 4 as providing a complete solution to the problem analyzed in the preceding sections. This, however, is not the case. In order to calculate B^*, Congress must in general know both V and C and calculate q^* and x^*. In this case Congress would not need to delegate decision-making authority to the military. It could simply instruct the military to choose q^*. That is, the rationale for having Congress delegate the quality choice to the military was the military's greater expertise in evaluating the potential costs and benefits of various design choices. However, the use of fixed budget levels provides a first-best outcome only when Congress is fully expert itself. Thus precommitment to fixed budget levels is not likely to be a good solution when Congress is very poorly informed about possible design choices.

Proposition 4 does suggest, however, a more limited sense in which delegated design choice with fixed-budget levels may be useful. Suppose that Congress is relatively well informed about the possible design choices for some project and thus can directly calculate q^*, x^*, and B^*. The following two strategies for Congress yield identical outcomes. First, Congress could delegate no decision-making authority and simply order the military to produce the quality q^*. In the procurement phase Congress would then purchase x^* units. The second possible strategy for Congress would be to tell the military that the project budget will be B^* and to delegate the choice of q and x.

Even though both strategies yield identical outcomes it may be that the second strategy of using delegation together with a fixed budget would be preferable from a transactions-cost point of view. Telling the military that its budget is B^* is a very simple, objectively verifiable instruction. However, directly ordering the military to choose the quality q^* may involve a very complex set of instructions. This is because no simple objectively verifiable scalar measure of quality is likely to exist. To provide an objectively verifiable description of the design choice yielding q^* could be quite complex and difficult. Recall that the military will in general prefer to choose some other quality level than q^*, so any design description which leaves leeway in the military's choice is likely to be purposefully misinterpreted.

There is also a second major problem with the fixed-budget approach suggested by Proposition 4. For a major new program the design decision may be made five years or more before procurement begins and procurement may continue over a 10- or 20-year period. It is hard to believe that Congress would have the ability to precommit 10 years in advance to anything. More importantly, however, many factors in the environment will change between the time the design is chosen and the time that procurement decisions are made. A first-best program should involve changing the amount procured in response to these changes in the environment. Thus, in a more realistic model, the first-best solution does not involve fixing a budget level at the time of the design choice. This issue will now be explored in a simple formal model.

---

[7] The military actually only chooses q. Then Congress chooses x so that the program costs B dollars. Formally, however, this is equivalent to having the military choose q and x subject to the constraint that the program cost at most B dollars. Thus the military will be viewed as choosing q and x.

Let $\theta \in \Theta$ represent environmental factors which are unknown at the time of the design decision. Assume θ is distributed according to the density $f(\theta)$. Both the military value and cost of a new system could potentially depend on θ. Thus these are now written, respectively, as $V(q,x,\theta)$ and $C(q,x,\theta)$.

The sequence of decisions is now as follows. At the time of the design decision θ is unknown. After the design decision is made θ becomes known and then the output decision is made. Therefore a program is now an ordered pair $(q,\psi(\theta))$, where q denotes the quality choice and $\psi(\theta)$ denotes the output decision given θ.

A first-best program is defined as follows. Let $E\{\ \}$ denote the expectation operator over θ.

Definition 4: The program $(q^*,\psi^*(\theta))$ is first-best optimal if it satisfies

$$(23)\qquad \psi^*(\theta) \in \underset{x\geq 0}{\operatorname{argmax}}\, V(q^*,x,\theta) - C(q^*,x,\theta)$$

for every $\theta \in \Theta$ and

$$(24)\quad q^* \in \operatorname{argmax} E\{V(q,\psi^*(\theta),\theta) - C(q,\psi^*(\theta),\theta)\}.$$

Under the first-best solution, the budget level depends on the realization of θ. Let $B^*(\theta)$ denote the first-best budget rule for the optimal program $(q^*,\psi^*(\theta))$. It is defined by

$$(25)\qquad B^*(\theta) = C(q^*,\psi^*(\theta),\theta).$$

Using the same reasoning as in the proof of Proposition 4, it is clear that Congress could induce the military to choose q^* by precommitting to the fixed budget rule defined by (25). That is, Congress should commit to spending $B^*(\theta)$ dollars if θ occurs independent of the quality level chosen by the military. This will induce the military to choose q^* and then $\psi^*(\theta)$ units will be procured.

Thus in theory there still exists a method for Congress to induce the optimal quality decision. It simply commits to a fixed-budget *rule* instead of a fixed-budget *level*. However, Congress is much less likely to be able to precommit to such a *rule* for two related reasons. First, the space of possible environmental factors which might affect the social cost or benefit of a new weapons program five or ten years in the future is almost limitless in size. Thus describing the set of all major conceivable contingencies and the budget level for each one would in general be an impossible task. A second related problem is that many of the contingencies described by θ may be observable only to Congress or at least are difficult to describe in an objectively verifiable way. Thus it may be difficult for Congress to describe the rule it will follow in such a way that whether or not it follows the rule will be objectively verifiable and noncontroversial.

The conclusion of this section is therefore that the organizational device of precommitting to fixed budgets will be of limited usefulness for major new weapons systems where design choices involve new technologies and there are long time lags between the design decision and actual production. Fixed budgets may play a more useful limited role in shorter-run projects involving better understood technological choices.

IV. Overlapping Service Jurisdictions

Overlapping service jurisdictions are said to occur when two or more of the three military services (the army, navy, and air force) share and compete for responsibility for the same military function or mission. Traditional military analysts have noted these overlaps and cited them as examples of inefficient organization within the military.

There is undoubtedly some truth to this point of view. Having two services each perform overlapping portions of the same mission and compete for the right to perform new related functions may well produce outcomes that seem inefficient relative to the full-information first-best standard. Thus in a world with no asymmetric information or incentive issues complicating the problem of organizational design, there may be little role for overlapping service jurisdictions.

However, overlapping jurisdictions may play a more useful role in a world where the military services do not necessarily have identical preferences as Congress and where Congress is not as fully informed as the military. In this case Congress (or the DoD) can use rivalry between competing services to improve the performance of each service. Specifically, it is straightforward to show that in this paper's model, the first-best program can be achieved by having two services compete for the right to operate a given program. See Rogerson (1988) for details.

V. Application to the Theory of Organization of the Firm

The model of this paper can also be interpreted as applying to the relationship between the central management of a firm and the manager of a particular project or division within the firm. Central management delegates selection of many aspects of the operation of a project or a division to the manager directly in charge of it because it does not have the time or expertise to make all of these decisions itself. However it retains final budgetary authority for the project or division.

In terms of the notation of the model, q is the design parameter for the project which central management delegates to the project manager. The project manager reports his decision of q to central management and then central management decides the "level" to operate the project at, denoted by x. The revenues to the company from the project are given by $V(q,x)$ and the costs of the project are given by $C(q,x)$. The parameter q can still be interpreted as a quality choice since higher values of q raise both revenues and costs. The parameter x does not necessarily have a simple physical interpretation such as "units purchased." In terms of realistically interpreting the model it might be better to suppress the variable x and think of central management as simply choosing a budget level given the project manager's design announcement of q. When the project manager announces a quality level $\hat{q}$ for the project he can be viewed as informing management that any revenue-cost pair from the set

$$\{(V(\hat{q},x),C(\hat{q},x)) \colon x \geq 0\} \tag{26}$$

is possible. Central management then chooses a budget level for the project and a corresponding revenue. Higher values of x correspond to the choice of a higher budget and a higher revenue stream. Thus x can be thought of as a choice of what "level" to run the project at. In some cases x may be units purchased. For example the project could be an advertising campaign and x could be minutes of airtime purchased. In other cases x may be some aspect of quality which central management reserves to decide for itself. However, in general central management should be thought of as making a budget choice when confronted with the set (26). The variable x may not have a specific real interpretation but is simply a technical convenience for describing central management's choice of a budget.

The remaining question to investigate is the nature of the project manager's preferences. As with the Congress-military interpretation, the key factor determining this is whether V and C can be easily measured in an objectively verifiable way. If they are objectively measurable, central management should be able to use incentive contracts to directly induce project management preferences defined by (5). However, when they are not objectively measurable, the project manager's personal preferences are likely to play a role in determining his actions. As argued in Section I, preferences defined by (4) are then as idealistic as could be realistically hoped for. In the Congress-military interpretation, V was clearly not objectively measurable. In the interpretation of this section, V and C may or may not be objectively measurable. The conclusions of this paper apply to cases where they are not measurable.

The key factor determining whether V and C can be easily measured in how separable the project is from the rest of the firm's operations. Suppose for example that the project is for the firm to enter a new market and to build a totally separate organization to market and produce it. Then the project's revenues and costs will accrue separately from those of the rest of the firm and are thus easy to objectively measure. However, now suppose that the output of the project will simply be an input for the production process of the entire firm. For example, the project might be an advertising campaign or creation of a new computing system. No objectively verifiable measure of the marginal contribution of the project to the firm's revenues is likely to exist. There may also be problems with objectively measuring $C(q, x)$ if the project is purchasing some of its inputs from other divisions of the firm.

VI. Conclusion

This paper can be viewed as making a general point about organizational design. When the divisions of an organization are not totally separate entities but rather contribute inputs to a complex joint productive effort, central management cannot really expect its subordinates to attempt to perform the calculation of how funds should be allocated across the divisions. This is the job of central management. Rather the best that central management can realistically hope for is that its subordinates in good faith attempt to produce the most effective program they can. If the central management could calculate the first-best allocation of budget levels across its subordinate divisions this would yield the first-best outcome. However, the limited calculation ability of the central authority means that the central authority may have to delegate design decisions to its subordinates and choose budget levels given the design decisions. This results in nonoptimal outcomes. In particular, if a division can increase its budget by suggesting a higher quality, then quality will be too high. This occurs even though subordinates derive no consumption value from quality and are in some sense acting in good faith to benefit the central authority.

REFERENCES

Baron, David and Besanko, David, "Monitoring, Moral Hazard, Asymmetric Information, and Risk Sharing in Procurement Contracting," *Rand Journal of Economics*, Winter 1987, *18*, 509–32.

Gansler, Jacques, *The Defense Industry*, Cambridge, MA: MIT Press, 1980.

Lewis, Tracy, "Budget Competition in Project Funding," mimeo., 1986.

McAfee, Preston and McMillan, John, "Bidding for Contracts: A Principal Agent Analysis," *Rand Journal of Economics*, Autumn 1986, *17*, 326–38.

Peck, Merton and Scherer, Frederic, *The Weapons Acquisition Process*, Boston: Harvard University Graduate School of Business Administration, 1962.

Riordan, Michael and Sappington, David, "Designing Procurement Contracts," mimeo., 1986.

Rogerson, William, "Quality vs. Quantity in Military Procurement: An Organizational Theory of Decision Bias," Northwestern Centre for Math Studies in Economics and Management Science Working Paper No. 771, 1988.

Scherer, Frederic, *The Weapons Acquistion Process: Economic Incentives*, Boston: Harvard University Graduate School of Business Administration, 1964.

Stubbing, Richard, *The Defense Game*, New York: Harper & Row, 1986.

Tirole, Jean, "Procurement and Renegotiation," *Journal of Political Economy*, April 1986, *94*, 235–59.

Tobias, Sheila, Goudinoff, Peter, Leader, Stefan and Leader, Sheleh, *The People's Guide to National Defense*, New York: William Morrow, 1982.

A Theory of Predation Based on Agency Problems in Financial Contracting

By PATRICK BOLTON AND DAVID S. SCHARFSTEIN*

By committing to terminate funding if a firm's performance is poor, investors can mitigate managerial incentive problems. These optimal financial constraints, however, encourage rivals to ensure that a firm's performance is poor; this raises the chance that the financial constraints become binding and induce exit. We analyze the optimal financial contract in light of this predatory threat. The optimal contract balances the benefits of deterring predation by relaxing financial constraints against the cost of exacerbating incentive problems. (JEL 610)

In this paper, we present a theory of predation based on agency problems in financial contracting. Our work is closest in spirit to the "long-purse" (or "deep-pockets") theory of predation, in which cash-rich firms drive their financially constrained competitors out of business by reducing their rivals' cash flow.[1] Although the existing theory is suggestive, it begs important questions. Why are firms financially constrained? And, even if firms are financially constrained, why don't creditors lift these constraints under the threat of predation?

We attempt to answer these questions. In Section I, we present a model (which is of independent interest) in which financial constraints emerge endogenously as a way of mitigating incentive problems. We argue that the commitment to terminate a firm's funding if its performance is poor ensures that the firm does not divert resources to itself at the expense of investors. This termination threat, however, is costly in a competitive environment. Rival firms then have an incentive to ensure that the firm's performance is indeed poor. This increases the likelihood that investors cut off funding, and induces premature exit.

In Section II, we analyze the optimal contract when firms and investors take this cost into account. In general, the optimal response to predation is to lower the sensitivity of the refinancing decision to firm performance. There are two ways of doing this. One is to increase the likelihood that the firm is refinanced if it performs poorly; the other is to lower the likelihood that the firm remains in operation even if it performs well. Both strategies reduce the benefit of predation by lowering the effect of predation on the likelihood of exit. We identify conditions under which each of these strategies is optimal.

There is a tradeoff between deterring predation and mitigating incentive problems; reducing the sensitivity of the refinancing decision discourages predation, but exacerbates the incentive problem. Depending on the importance of the incentive problem relative to the predation threat, the equilibrium optimal contract may or may not deter predation.

We are by no means the first to present a theory of rational predation. In the existing models of rational predation,[2] one firm tries

---

*Department of Economics, Harvard University, Cambridge, MA 02138 and Sloan School of Management, MIT, Cambridge, MA 02139. This paper is a revised version of "Agency Problems, Financial Contracting, and Predation." For helpful comments, we thank Drew Fudenberg, Oliver Hart, Patrick Rey, Julio Rotemberg, Jean Tirole, the anonymous referees, and seminar participants at Harvard, MIT, and the European Conference on Information Economics in Madrid.

[1] See, for example, John McGee (1958), Lester Telser (1966), Jean-Pierre Benoit (1984), and Jean Tirole (1988).

[2] See for example, Steven Salop and Carl Shapiro (1982), David Scharfstein (1984), and Garth Saloner (1987). These papers draw much from the early work of Paul Milgrom and John Roberts (1982) on rational limit-pricing.

to convince its rivals that it would be unprofitable to remain in the industry; predation changes rivals' beliefs about industry demand or the predator's costs. In our model, there is common knowledge that production in each period is a positive net present value investment. Thus, predation need not be effective by changing rivals' beliefs, but rather by adversely affecting the agency relationship between the firm and its creditors. Drew Fudenberg and Jean Tirole (1986) have also argued that agency problems between creditors and the firm can result in financial constraints that induce predatory behavior by rivals. Their model differs from ours in that they consider only a sequence of one-period contracts and do not consider optimal responses to predation.

Our paper is also related to the recent work on the interaction between product-market competition and the capital market. James Brander and Tracy Lewis (1986) and Vojislav Maksimovic (1986) are among the earliest papers. They point out that because equity holders receive only the residual above a fixed debt obligation, the marginal production incentives of managers who maximize the value of equity depends on the debt-equity ratio. Therefore, investors can use capital structure to induce managers to compete more agressively, in the process affecting product-market equilibrium. There are two drawbacks of their work. First, it restricts attention to a subset of feasible financial instruments; under a broader set of instruments, product-market equilibrium would be very different. Our paper, in contrast, derives the set of feasible contracts from first principles and analyzes optimal contracts within that set. Second, in these papers, financial structure plays no role other than through its effect on product-market strategy. In our analysis, financial policy also affects agency problems within the firm.

These papers are similar in spirit to the work of John Vickers (1985) and Chaim Fershtman and Kenneth Judd (1987a, b) who analyze the effect of managerial incentive contracting (rather than financial contracting) on product-market competition. Like Brander and Lewis's model, firms gain competitive advantage by altering managerial objectives. For example, by basing compensation on sales, shareholders can induce the manager to produce more output. This may have strategic value if firms compete in a Cournot environment. Unlike Brander and Lewis's model, but like ours, the latter Fershtman and Judd paper analyzes optimal contracts. These contracts serve the dual function of mitigating agency problems and affecting product-market competition.

On a formal level, our basic framework is similar to work by Roger Myerson (1982) and Michael Katz (1987). These papers consider incentive problems between a principal and an agent in which the agent's performance both influences and is influenced by other parties. Although Myerson's development is in an abstract principal-agent setting and Katz's main application is to bargaining, this framework seems particularly well-suited to analyze the interaction between product-market competition and the capital market.

An important implication of our approach is that financial structure affects firms' financing costs as well as their gross profitability. Information and incentive problems in the capital market can determine the structure of the product market. This is in contrast to most models, in which capital structure only affects financing costs.[3]

Finally, we note that the dynamic nature of our of financial contracting has some novel features. We assume that contracts cannot be made directly contingent on profits. To induce managers to pay out cash flows to investors, the firm is liquated when its payouts are low. Liquidation occurs although it is inefficient, and the threat of liquidation is credible. The optimal dynamic financial contracts that we describe resemble standard debt contracts in many ways.

I. Contracting Without Predation

There are two firms labeled A and B, who compete in periods 1 and 2. At the beginning of each period, both firms incur a fixed cost,

---

[3] This is also a feature of Robert Gertner, Robert Gibbons, and Scharfstein (1988). In that paper capital structure decisions can convey information to both the capital and product markets.

F. The firms differ with respect to how they finance this cost. Firm A has a "deep pocket," a stock of internally generated funds which it can use to finance this cost. In contrast, firm B has a "shallow pocket"; it must raise all funds from the capital market.

The first step in our analysis is to characterize the contractual relationship between firm B and its sources of capital. We assume that there is one investor who makes a take-it-or-leave-it contract offer to firm B at the initial date 0, which firm B accepts if the contract provides nonnegative expected value. The assumption that the investor rather than the firm has all the bargaining power may seem unrealistic. This is particularly so for a firm issuing public debt or equity in a well-functioning capital market with many competing investors. However, young companies requiring venture capital, or older ones placing private debt or equity, are likely to bargain with investors. In reality, neither side has all the bargaining power; our assumption simply sharpens our results without affecting their essential character. Indeed, it will become clear below that the termination threat must be part of any feasible contract regardless of the competitive structure of the capital market.

Firm B's gross profit (before financing costs) in each period is either π_1 or π_2, where $\pi_1 < \pi_2$. At the beginning of each period, all players believe that $\pi = \pi_1$ with probability θ. Thus, we are assuming that profits are independently distributed across periods. We make this assumption to distinguish our results from models of predation (based on the limit-pricing model of Milgrom and Roberts, 1982) in which the incumbent firm tries to convince the entrant that it would be unprofitable to remain in the industry. These models rest on the assumption that the entrant's profits are positively serially correlated. As we argue below, we would strengthen our results by assuming that profits are positively correlated.

For simplicity, we assume that the discount rate is zero. We also assume that $\pi_1 < F$: with positive probability the investment loses money. Later we analyze the model under the assumption that $\pi_1 > F$ and argue that in that case as well the termination threat is valuable (although the analysis raises some other issues). In both cases, the expected net present value of the investment is positive:

$$\bar{\pi} \equiv \theta\pi_1 + (1-\theta)\pi_2 > F.$$

The agency problem we analyze stems from the impossibility of making financial contracts explicitly contingent on realized profit. There are two alternative interpretations of this assumption. One is that at the end of each period, the firm privately observes profit. The other is that profit is observable but not verifiable; although both the firm and investor can observe profit, the courts cannot, and hence these parties cannot write an enforceable profit-contingent contract.[4]

We do assume that the investor can force the firm pay out a minimum of π_1 in each period. If the courts know that this is the minimum possible profit, then a contract of this form is feasible. This assumption amounts to the claim that π_1 is the verifiable component of profit and the residual $\pi_2 - \pi_1$ is the non-verifiable component.

There are at least three reasons for assuming that contracts cannot be made fully contingent on realized profit. First, it is often difficult to judge whether particular expenses are necessary; what look like justifiable expenses may really be managerial perquisites with no productive value. Thus, there is scope for managers to divert resources away from investors to themselves. A second, related reason is that the firm might be affiliated with another firm, thus providing some flexibility in the joint allocation of costs and revenues. Finally, from a methodological perspective, the assumption that profits are not observable generates simple and intuitive results that generalize to a wide variety of

[4] In many situations, the latter interpretation is more plausible; an investor is often closely involved in the firm's operations, whereas the courts are not. Irregular accounting practices can make it difficult for outside parties to know the firm's true profitability. Although these assumptions have the same implications in the basic model we analyze, they will have different implications if renegotiation is possible. We discuss this in more detail in Section II, Part B.

realistic agency models. One such model is discussed in Section II, Part B.[5]

In a one-period model, the investor would not invest in the firm. To see this, let R_i be the transfer from the firm to the investor at date 1 if the manager reports that profit is π_i, $i=1,2$. Assuming limited liability protects the firm and its managers, and that the firm has no other assets, R_i can be no greater than π_i. Clearly, it will report the profit level that minimizes financing costs. Since $R_i \leq \pi_i$, at date 1 the investor can receive at most $\pi_1 < F$ and hence would always lose money. If instead the relationship lasts for two periods, the investor can control whether the firm receives financing in the second period. The investor can threaten to cut off funding in the second period if the firm defaults in the first. This threat induces the firm to pay more than π_1 in the first period. Note that this threat is credible; since $\pi_1 - F < 0$, no investor wishes to finance the firm in the second period.

Formally, we analyze the contract-design problem as a direct revelation game, in which the terms of the contract are based on the firm's report of its profit. In particular, suppose the investor gives the firm F dollars at date 0 to fund first-period production. As in the above one-period model, let R_i be the transfer at date 1 if the firm reports profits of π_i in the first period. Let $\beta_i \in [0,1]$ be the probability that the investor gives the firm F dollars at date 1 to fund second-period production if the firm reports π_i in the first period.[6] We assume that without this second-round financing the firm lacks the necessary funds to operate in the second period.[7] Finally, let R_{ij} be the transfer from the firm to the investor at date 2 if the first-period report is π_i and the second-period report is π_j.

It is clear from the argument presented above for the one-period model that $R_{i1} = R_{i2}$; the second-period transfer cannot depend on second-period profit because the firm would always report the profit level corresponding to the lower transfer. Thus, let R^i be the second-period transfer if the first-period reported profit is π_i. It follows from the limited liability assumption that $R^i \leq \pi_i - R_i + \pi_1$; the second-period transfer cannot exceed the surplus cash from the first period, $\pi_i - R_i$, plus the minimum profit in the second period, π_1.[8]

The optimal contract maximizes the expected profits of the investor subject to the following constraints: (1) the firm truthfully reveals its profit at dates 1 and 2 (incentive compatibility); (2) the contract does not violate limited liability; (3) the firm opts to sign the contract at date 0 (individual rationality). Formally, the problem is the following:

$$\underset{\{\beta_i,\, R_i,\, R^i\}}{\text{Maximize}} - F + \theta\left[R_1 + \beta_1(R^1 - F)\right] + (1-\theta)\left[R_2 + \beta_2(R^2 - F)\right],$$

subject to

$$(1) \qquad \pi_2 - R_2 + \beta_2(\bar{\pi} - R^2) \geq \pi_2 - R_1 + \beta_1(\bar{\pi} - R^1);$$

$$(2) \qquad \pi_i \geq R_i,$$

$$(2') \qquad \pi_i - R_i + \pi_1 \geq R^i, \qquad i=1,2;$$

$$(3) \qquad \theta\left[\pi_1 - R_1 + \beta_1(\bar{\pi} - R^1)\right] + (1-\theta)\left[\pi_2 - R_2 + \beta_2(\bar{\pi} - R^2)\right] \geq 0.$$

---

[5] Robert Townsend (1979) and Douglas Gale and Martin Hellwig (1985) present models in which investors and firms can write profit-contingent contracts for some finite cost; this contrasts with our assumption of infinite costs. Below, we discuss the relationship between these models and ours.

[6] For now, we are assuming that there exists an enforceable randomization scheme. We discuss this assumption in greater detail below.

[7] One can show that this amounts to assuming that $\pi_2 - \pi_1 < F$.

[8] Implicit in this formulation is the assumption that the firm must keep profits (net of transfers to the investor) in the firm between dates 1 and 2, but that at the end of period 2 any profit left over can be consumed by the entrepreneur. This is consistent with the assumption that profits cannot be observed.

The incentive-compatibility constraint (1) ensures that when profit is high the firm does not report that profit is low. If profit is π_2, the firm receives some surplus in the first period if it reports π_1 since $R_1 \leq \pi_1 < \pi_2$; however, by setting $\beta_1 < \beta_2$ the investor makes it costly for the firm to report π_1, since the firm generally receives surplus in the second period.

We have omitted the incentive-compatibility constraint ensuring that the firm reports π_1 rather than π_2. We demonstrate later that this constraint is not binding. Note also that the limited-liability constraints (2) and (2′) imply that the individual-rationality constraint (3) is not binding.

The following two lemmas, which we prove in the Appendix, simplify analysis of the optimal contract.

LEMMA 1: *The incentive compatibility constraint* (1) *is binding at an optimum.*

LEMMA 2: *There exists an optimal contract in which second-period transfers, R^1 and R^2, equal π_1.*

Lemma 1 is a typical feature of contracting problems. Lemma 2 establishes that the investor can receive at most π_1 from the firm in the second period because there is no termination threat at that time.

These two lemmas simplify the maximization problem to

$$\text{(4)} \quad \text{Maximize } -F+R_1 + \beta_2(1-\theta)(\bar{\pi}-F) - \beta_1[\theta F+(1-\theta)\bar{\pi}-\pi_1],$$

subject to the limited-liability constraint, $\pi_i \geq R_i$, $i=1,2$.

Let $\{R_1^*, \beta_1^*, R_2^*, \beta_2^*\}$ denote the optimal contract. It follows immediately that $R_1^* = \pi_1$ and $\beta_2^* = 1$. Moreover, because both F and $\bar{\pi}$ exceed π_1, the last bracketed term is positive. Thus, $\beta_1^* = 0$. It then follows from the incentive-compatibility constraint (1) that $R_2^* = \bar{\pi}$. Finally, this contract satisfies the limited-liability constraints and (given that $R_2^* = \bar{\pi} > \pi_1$) the omitted incentive constraint.[9]

There are two reasons why the investor cuts off funding if the firm reports low profits. First, the investor avoids losing $F-\pi_1$ in the second period. Second, it induces the firm to report profits truthfully, enabling the investor to extract more surplus from the firm in the first period. To see this, note that the incentive constraint implies

$$\text{(5)} \quad R_2^* = \pi_1 + (1-\beta_1)(\bar{\pi}-\pi_1).$$

The term $\bar{\pi}-\pi_1$, is the firm's expected surplus in the second period given it operates then. By reporting π_1 rather than π_2, the firm reduces by $(1-\beta_1)$ the probability that it receives this surplus. A marginal reduction in β_1 therefore lowers by $\bar{\pi}-\pi_1$ the expected value of reporting π_1. Hence, it increases by $\bar{\pi}-\pi_1$ the amount the investor can require the firm to pay when it reports profit of π_2.

Finally, we must determine the conditions under which the investor earns nonnegative profit. Given the optimal contract, the investor's expected profits are $\pi_1 - F + (1-\theta)(\bar{\pi}-F)$. Thus, for the investor to invest at date 0, F can be no greater than $\bar{\pi}-(\bar{\pi}-\pi_1)/(2-\theta)$. As a result, some positive net present value projects may not be funded.

We summarize these results in the following proposition.

PROPOSITION 1: *The investor invests at date 0, if and only if $F < \bar{\pi}-(\bar{\pi}-\pi_1)/(2-\theta)$. In this case, $R_1^* = \pi_1$, $\beta_1^* = 0$, $R_2^* = \bar{\pi}$, $\beta_2^* = 1$; the firm operates in the second period if and only if its first-period profits are π_2.*

The proposition implies that there is an *ex post* inefficiency; the firm is liquidated when first-period profit is π_1 even though

---

[9] Note that the firm weakly prefers to announce the true profit, when first-period profit is π_1. If the firm reports π_2, it is unable to make its first-period payment. In this case, the investor is paid π_1 and does not refinance the firm. The firm is then indifferent between the profit reports, in which case we assume that it reports its true profit.

$\bar{\pi} > F$ and it is efficient to operate.[10] It is natural to ask whether, at date 1, after first-period profit of π_1 is realized, the two parties wish to tear up the original contract and renegotiate a mutually beneficial arrangement. Note, however, that although it is efficient to produce, the most the investor can receive from the firm is $\pi_1 < F$. Thus, it is impossible to negotiate around the contractually specified inefficiency and no other investor would be willing to lend money. *Our results therefore do not depend on the assumption that renegotiation is impossible or that other investors are irrational.*[11]

A. *Discussion*

The main point of the analysis is that a firm's performance affects its financing costs *and* its access to capital. This result is quite general and captures an important feature of corporate-financing arrangements. For example, in venture-capital financing the venture capitalist rarely provides the entrepreneur with enough capital up front to see a new product from its early test-marketing stage to full-scale production. (See William Sahlman, 1986) Instead, typical venture-financing arrangements take the form of "staged capital commitment." Initially, the venture capitalist provides enough money to finance the firm's start-up needs like research and product development. Conditional on the firm's performance in this early stage, the venture capitalist may provide further financing to fund test-marketing, and then full-scale production.

There are at least two reasons why such contracts are used. First, they mitigate adverse selection problems. Entrepreneurs who have confidence in the venture accept contracts of this form more willingly because they know that when they return for more funding it will be at favorable terms. This point is similar to Mark Flannery's (1986) explanation of short-term debt and Benjamin Hermalin's (1986) argument that more able workers will sign short-term contracts to signal their ability. Second, staged financing arrangements reduce incentive problems between entrepreneurs and financiers. Requiring the firm to return to the venture capitalist for further funding limits the extent to which management can pursue its own interests (like consuming excess cash) at the expense of the venture capitalist. Our model formalizes this second benefit of staged capital commitment.

The model applies to more than just venture capital. Any disbursement of corporate funds through, for example, debt payments, dividends, or share repurchases, increases the chance that the firm will be unable to finance investment internally and must return to the capital market for further financing. And, as argued above, the commitment to go back to the capital market can increase value either through the information it conveys or its effect on managerial incentives. Michael Jensen (1986) has made a similar point: forcing managers to pay out cash prevents them from spending free cash flow on unprofitable investment projects.

Joseph Stiglitz and Andrew Weiss (1983) have also argued that the termination threat is an effective incentive device. Their analysis, however, differs from ours in two ways: the contracts they consider are not optimal; and the incentive problem concerns the choice of project riskiness rather than the observability of profits.

Finally, the agency problem we analyzed is related to the one-period models of Townsend (1979) and Gale and Hellwig (1985). In these models, the investor has a costly inspection technology that enables him to make payments contingent on profits. The optimal contract specifies that if the firm reports low enough profits, then the investor inspects and confiscates all of the firm's profits. Thus, inspection in these models plays the same role as the termination threat.[12] There is one important difference, however. In the inspection models it is never optimal for the investor to inspect once the firm has reported low profits even though the contract calls for him to do so; at this point, the investor knows the firm's profits and need not inspect. Thus, the inspection threat is not credible. In contrast, the termination threat in our model is credible; since $\pi_1 < F$, the investor always prefers to cut off funds when the firm's profits are low.[13]

---

[10] This result is reminiscent of Townsend (1982) where an inefficiency in the second period facilitates trade in the first period. Like our model, trade cannot be supported if there is only one period.

[11] A more subtle set of questions arise if the firm reports π_1, when profits are really π_2 and then tries to renegotiate the contract. We turn to this question in the next section.

B. *Extensions*

Other Agency Problems. Our model focuses on a particular agency problem, which enables us to make our point in the simplest possible way. We believe that the termination threat is useful for a wide variety of agency problems. The following example exhibits how this basic idea extends to the familiar effort-elicitation model of agency. This model also shows that our results do not depend on the assumption that profits are privately observed.

Suppose there are two periods of production and that in each period the manager can "work" or "shirk." By working, the risk neutral manager increases the probability that profit is π_2, but he incurs a utility cost. If the manager has limited wealth or is protected by limited liability, investors cannot sell him the entire firm (which is otherwise the optimal solution to the moral-hazard problem when the manager is risk neutral). This implies that if there is only one period of production, the manager must receive some expected surplus to induce him to work. (See Sappington, 1983, for a result along these lines.) This is analogous to the result in the one-period model considered above in which the firm receives an expected surplus of $\bar{\pi} - \pi_1$. The manager, therefore, bears a cost if the firm is not refinanced. Thus, the threat of not refinancing the firm raises the cost of shirking; the investor can then induce greater effort at lower cost.

Correlation in Profits. In this model, profits are independently distributed across periods. Thus, unlike many multiperiod agency models, the principal (investor) learns nothing about the agent's (firm's) profitability over time. We can extend the model to the case where profits are positively serially correlated. In fact, this strengthens our results.

Let $E(\pi|\pi_i)$ be expected second-period profits conditional on first-period profits, π_i. With positive serial correlation, $E(\pi|\pi_2) > \bar{\pi}$ and $E(\pi|\pi_2) > E(\pi|\pi_1)$. It is straightforward to establish that the optimal contract sets $\beta_1^* = 0$, $\beta_2^* = 1$, $R_1^* = \pi_1$, and $R_2^* = E(\pi|\pi_2)$. Since the firm's expected surplus in period 2 is greater when first-period profit is π_2, it loses more if it is not refinanced. This reduces the manager's incentive to underreport profits and enables the investor to extract more rent from the firm.[14]

Capacity Expansion. So far we have interpreted β_i as the probability of refinancing. Alternatively, we can interpret β_i as a capacity-expansion parameter; the investor commits to a staged capital-expansion plan contingent on the firm's first-period performance. By this we mean that if profits are π_i, the investor gives the firm enough money to increase capacity by an amount $\beta_i F$. We assume this increases expected profits by $\beta_i \bar{\pi}$.[15] Under this interpretation, we drop the constraint $\beta_i \in [0,1]$, and suppose that β_i lies in some interval $[\underline{\beta}, \bar{\beta}]$. These assumptions preserve the basic structure of our model. Thus, the investor sets $\beta_1 < \beta_2$ to mitigate incentive problems.

---

[12] In our model, if inspection costs are finite, one can show that the investor would never simultaneously use both inspection and the refinancing threat. So if inspection costs are high enough, only the refinancing threat is used.

[13] We are grateful to Julio Rotemberg for pointing out this difference between the models.

[14] If the firm's profits are negatively correlated over time, it loses less if it is not refinanced and it is more difficult for the investor to extract rent from the firm. However, situations in which profits are negatively correlated over time seem rather implausible.

[15] The assumption that expected profits are linear in β is strong, but it could be relaxed without much difficulty.

Renegotiation. The model assumes that renegotiation is not feasible. Suppose instead that after profits are realized the firm and investor can renegotiate mutually beneficial changes in the contract. Further, suppose that profits are observable to both parties (although not verifiable).

If profits are π_2 but the firm reports π_1, under the terms of the contract the firm is not supposed to receive any further financing. However, the firm has leftover cash of $\pi_2 - \pi_1$. Thus, if $\pi_2 - \pi_1 > F$, the firm can finance second-period investment with its own funds. In this case, it is impossible to induce the firm to pay more than π_1 in the first period and thus no investor would finance investment.[16]

This result rests on the assumption that investors cannot enforce a covenant restricting further investment by the firm; however, if investment is verifiable, such a covenant is feasible. Thus, suppose the investor and firm agree on this covenant at date 0 and suppose the firm decides to report π_1 when its profit is π_2. Without permission from the investor, the firm is prohibited from further investment and is forced to liquidate.

Liquidation, however, is inefficient and we would expect the parties to renegotiate around this covenant. Thus, whether the firm is willing to deviate (by reporting π_1 rather than π_2) depends on the outcome of the renegotiation process. If the firm has none of the bargaining power, the most it stands to gain from deviating is $\pi_2 - \pi_1$, which is exactly what it would get if it did not deviate. Since the firm is indifferent between the two alternatives, we can assume it would report profits truthfully. If, instead, the firm has some of the bargaining power during renegotiation, it can extract some of the efficiency gains and thus will earn more than $\pi_2 - \pi_1$ from deviating. Thus, the contract is not "renegotiation-proof" and not incentive compatible.

A renegotiation-proof can be designed, however. Suppose that the firm stands to gain a fraction, α, of the efficiency gain from investing, $\bar{\pi} - F$. Then instead of requiring the firm to pay $\bar{\pi}$ if it reports π_2, a renegotiation proof contract requires the firm to pay $\bar{\pi} - \alpha(\bar{\pi} - F)$; this makes the firm indifferent between reporting π_2 truthfully and reporting π_1 and then renegotiating.

The more bargaining power the investor has (the greater is α), the less the investor can require the firm to pay when it reports profit of π_2. If α is large enough so that the firm has most of the bargaining power then the renegotiation-proof payment is so low that the investor cannot cover his financing costs. For example, if α equals one, the firm pays F when profit is π_2 and π_1 when profit is π_1; these payments are not enough to cover the investor's costs. In this case, the possibility of renegotiation at date 1 drives out investment at date 0. There is, however, a wide range of parameter values for which the threat of renegotiation does not affect investment behavior.

The same analysis essentially applies when $\pi_2 - \pi_1 < F$. The only difference is that in this case the firm need not write a covenant against further investment since the firm must return to the investor to raise additional funds. The investor may be willing to lend because the firm has collateral of $\pi_2 - \pi_1$. By reducing the required payment when profit is π_2, the investor can ensure that the contract is renegotiation proof. It may, also, be efficient to restrict the firm from borrowing funds elsewhere because such borrowing induces competition among creditors and transfers some of the bargaining power to the firm.

Other Extensions. The results are easily generalized to a continuum of profit levels. One can show that if first-period profits are greater than $\bar{\pi}$, $\beta = 1$ and the firm pays back $\bar{\pi}$ in the first period. If profits are below $\bar{\pi}$ the firm pays back all of its first-period profits and it is refinanced with some probability between zero and one; the greater the firm's profits the greater the likelihood it is refinanced. In many ways, this contract resembles a debt contract: the firm is supposed to pay back $\bar{\pi}$; if it does not, all of its profits are paid over to the creditor and there is some chance that the firm is liquidated.

---

[16] Note that if profit is indeed π_1 there are no excess funds so that the firm cannot invest in violation of the prohibition.

One can also extend the model to assume that there is competition among investors at date 0 so that effectively the firm has all the bargaining power initially. In this case one can show that the contract involves $\beta_2^* = 1$ and $1 > \beta_1^* > 0$. The termination threat will still be used, but to a lesser extent and the contract will be more efficient.

II. Predation and the Optimal Contract

In this section we model explicitly the interaction between the firm's financial policy and product-market competition. To begin, suppose the investor and firm B ignore the existence of firm A when designing an optimal financial contract; they assume that (stochastic) profits are exogenous. In this case, the financial contract is as described above. But, this makes it attractive for firm A to prey. If firm A can lower firm B's expected first-period profit (say, by reducing its price or increasing its advertising), then it can increase the probability that firm B exits. Firm A will do so if the costs of taking such actions are less than the expected benefits of becoming a monopolist.

To formalize these ideas, we model predation as follows: for a cost $c > 0$, firm A can increase from θ to μ the probability that firm B earns low profit, π_1, in period 1. If firm B exits, firm A becomes a monopolist and its second-period expected profits are π^m. If, instead, firm B remains in the market, firm A's expected profits are π^d. Thus, given a contract in which the pair $(\beta_1, \beta_2) = (0,1)$, the expected benefits of predation are $(\mu - \theta)(\pi^m - \pi^d)$. It preys provided $(\mu - \theta)(\pi^m - \pi^d) > c$, or defining $\Delta \equiv c/[(\mu - \theta)(\pi^m - \pi^d)]$, if $\Delta < 1$. If it does prey, the investor's expected profits are $\pi_1 - F + (1 - \mu)(\bar{\pi} - F)$.

More generally, given any financial contract of firm B, firm A preys if $(\beta_2 - \beta_1)(\mu - \theta)(\pi^m - \pi^d) > c$ or $(\beta_2 - \beta_1) > \Delta$. Hence, the benefits of predation depend on firm B's financial contract. Note that when the investors of firm B ignore the possibility of predation, they maximize the benefit of predation to firm A, since $\beta_2 - \beta_1$ is largest when $\beta_2 = 1$ and $\beta_1 = 0$. *The contract that minimizes agency problems, maximizes the rival's incentive to prey.* To make the analysis interesting, we assume for the remainder of the paper that the parameters are such that if $\beta_2 = 1$ and $\beta_1 = 0$, it is optimal to prey, that is, $\Delta < 1$.

To analyze the effect of financial contracting on product-market equilibrium we need to make two further informational assumptions. First, we assume that the courts cannot observe firm A's predatory action. This is a reasonable assumption in light of the difficulties legal scholars and economists have encountered in defining predation. And, even if a reasonable definition of predation did exist, the information that the courts would need to use it could make enforcement unworkable. For example, the courts would need detailed knowledge of demand functions to know whether a firm's advertising and pricing policies were predatory.[17]

Given that the court cannot observe predation, firm B and the investors cannot make the contract contingent on the predatory action of firm A. Notice that we do allow firm B and its investors to observe firm A's predatory actions. This distinguishes our model from signaling (Milgrom and Roberts, 1982) and signal-jamming (Fudenberg and Tirole, 1986) models of predation which rely on the assumption that predation is not observable.

Our second informational assumption concerns the observability by firm A of the contract between firm B and its investors. If the predator can observe the contract, then the investor can use the contract to influence firm A's actions. By reducing the sensitivity of the refinancing decision to first-period profit, that is, reducing the difference between β_2 and β_1, the investor reduces the gains from predation. For small enough values of $(\beta_2 - \beta_1)$ he deters predation. He can do so by strengthening the commitment to refinance the firm, that is, increasing β_1. In the extreme, he can deter predation by setting $\beta_2 = \beta_1 = 1$. That is, the investor can give firm B a "deep pocket," a commitment of resources to finance investment in both periods. Alternatively, the investor can deter predation by refinancing the firm less often, that is, reducing β_2. We refer to this as a "shallow-pocket" strategy.

---

[17] See, for example, Paul Joskow and Alvin Klevorick (1979) for one attempt at defining predation and a discussion of the difficulties in doing so. See Scharfstein (1984) for a model which takes account of the costs of detecting predation.

In many cases it is reasonable to suppose that contracts are observable; for example, the Securities and Exchange Commission requires all publicly held firms to disclose information on their financial structure. For privately held companies, however, there is no disclosure requirement. It may be more reasonable to assume that for these firms financial contracts cannot be observed. We therefore consider the two cases of observable and unobservable contracts.[18]

A. *Observable Contracts*

If contracts are observable, the investor can ensure that firm A does not prey by writing a contract that satisfies the following "no-predation constraint":

$$(6) \quad (\beta_2 - \beta_1)(\mu - \theta)(\pi^m - \pi^d) \leq c, \text{ or}$$

$$(6') \qquad (\beta_2 - \beta_1) \leq \Delta.$$

That is, the investor can deter predation by reducing the sensitivity of the contract to firm B's performance.

Recall that our formulation assumes that there exists a public randomization technology enabling the investor to set $\beta_i \in (0,1)$. Without such a technology, the investor is restricted to deterministic schemes. This means the only feasible contract in which the firm enters must set $(\beta_1, \beta_2) = (0,1)$. Thus, one cannot deter predation in this case. This strengthens our point that predation can occur as an equilibrium phenomenon.

To determine the efficient contractual response to predation if randomization is feasible (or if we interpret β as a capacity-expansion parameter), we first analyze the optimal contract that deters predation. We then compare this contract to the optimal contract given predation. The investor chooses the contract with the higher payoff, provided it earns nonnegative profit.

The optimal predation-deterring contract solves the following program:

$$\max_{\{\beta_1, R_i\}} \; -F + \theta[R_1 + \beta_1(\pi_1 - F)]$$
$$+ (1-\theta)[R_2 + \beta_2(\pi_1 - F)],$$

subject to the incentive constraint (1), the limited-liability constraint (2), and the no-predation constraint (6′).

This maximization problem is identical to the problem analyzed in Section I except for the constraint (6′) which ensures that no predation occurs in the first period. At an optimum, this constraint is binding and $\beta_2 - \beta_1 = \Delta$; otherwise, the optimal solution would be $\beta_2 = 1$, $\beta_1 = 0$, firm A would prey, and the constraint would be violated. Observing that, as before, $R_1 = \pi_1$, the binding incentive constraint (1) becomes $R_2 = \Delta\bar{\pi} + (1-\Delta)\pi_1$. Substituting these equalities into the objective function, we reduce the maximization problem to

$$\max - F + \beta_1(\pi_1 - F) + (1-\theta)\Delta(\bar{\pi} - F).$$

It then follows that $\beta_1^* = 0$ and $\beta_2^* = \Delta$; *the investor optimally deters predation by lowering the probability that the firm is refinanced when its profits are high.* This deters predation because there is little incentive for firm A to pay a cost, c, to ensure that profits are low. But, why lower β_2 rather than increase β_1? A change in either of these two variables has the same effect on the no predation constraint and the incentive constraint. However, increasing β_1 is costly because it increases the probability that the investor loses $F - \pi_1$ in the second period, whereas lowering β_2 reduces this probability.

---

[18] For more discussion of the different implications of contract observability, see Katz (1987).

Finally, note that the expected profit from following this predation-deterring strategy is $\pi_1 - F + (1-\theta)\Delta(\bar{\pi} - F)$. Thus, conditional on entering, firm B chooses to deter predation provided $(1-\theta)\Delta > 1-\mu$. And, if $\pi_1 - F + \max\{(1-\theta)\Delta, 1-\mu\}(\bar{\pi} - F) > 0$, it will be profitable to enter. We summarize these results below.

PROPOSITION 2: *Firm B enters if and only if*

$$\pi_1 - F + \max\{(1-\theta)\Delta, 1-\mu\}(\bar{\pi} - F) \geq 0.$$

If B enters, and $(1-\theta)\Delta \geq 1-\mu$, *the optimal contract deters predation. In this case,* $\beta_1^* = 0$, $R_1^* = \pi_1$, $\beta_2^* = \Delta$, *and* $R_2^* = \Delta\bar{\pi} + (1-\Delta)\pi_1 < \bar{\pi}$. *If firm B enters and* $(1-\theta)\Delta < 1-\mu$, *the contract is as given in Proposition* 1 *and firm A preys.*

The striking result that the shallow-pocket strategy optimally deters predation depends on the assumption that $\pi_1 < F$. Suppose instead that $\pi_1 > F$. By setting $\beta_1 = 0$, the investor is able to extract more surplus from the firm in the first period, but he foregoes positive surplus of $\pi_1 - F$ in the second period. Provided π_1 is not too large, the optimal contract still sets $(\beta_1, \beta_2) = (0, 1)$.[19]

In the presence of a predatory threat, however, $\beta_2 - \beta_1$, must equal Δ to deter predation. But, here the investor earns positive profit in the second period. So rather than reduce β_2, β_1 should be increased; this increases the probability that the investor earns a profit of $\pi_1 - F$ in the second period. In this case, a deep pocket is the optimal response to the predatory threat.

Note, however, that if $\pi_1 > F$ it is inefficient for firm B to exit in the second period. Thus, if the contract calls for the firm to be liquidated in the second period when its first-period profits are low, the investor and the firm may be able to renegotiate a more efficient arrangement in which the firm remains in operation and the two parties split the surplus $\pi_1 - F$. If bargaining is efficient, firm B would never exit and firm A has no incentive to prey.

There are reasons to believe, however, that efficient renegotiation may not always occur. First, it can be in the investor's interest to ensure that such renegotiation is infeasible even though it encourages predation. By doing so, he may be able to extract more surplus from the firm in the first period. One way of committing not to renegotiate is to bring in other investors to finance the firm. If each has only a small stake and there are costs of negotiation, then none will have an incentive to renegotiate even though it would be efficient to do so if there were only one investor.[20] Second, our model assumes symmetric information about future profitability. A more realistic model would allow for the possibility of asymmetric information in which case bargaining is more likely to break down.

B. *Unobservable Contracts*

The assumption that contracts are observable may be inappropriate in some circumstances. There are no financial disclosure requirements for privately held firms, so it may be impossible for outsiders to observe a firm's contractual relationship with its creditors. Thus, we also investigate the case in which firm A cannot observe the contract signed by firm B and the investor. Instead, firm A must make a rational conjecture about the chosen contract.

When contracts are unobservable, it is as if the investor and the predator play a simultaneous move game.[21] (We can ignore firm B

---

[19] The precise condition, discussed in more detail in an earlier version of our paper, is that $\theta F + (1-\theta)\pi - F > 0$. This can be seen from inspection of (4) in the text.

[20] For example, small bondholders do not typically participate in financial renegotiations. Bankruptcy law recognizes this difficulty and provides a mechanism for facilitating renegotiation through the Chapter 11 reorganization process.

[21] Note, however, that we continue to assume that firm B signs the contract before firm A decides whether to prey. It might be argued that if the contract is signed first it would be in the interest of firm A to reveal its contract. But, as Katz points out in a related context, if the observed contract is not efficient, the two parties

because its actions follow trivially from the contract the investor chooses.) Firm A's strategy set is composed of two pure strategies: "prey," which we denote by P, and "do not prey," which we denote by NP. The investor's strategy set is essentially a choice of a pair $(\beta_1, \beta_2) \in [0,1]^2$. (We can ignore R_1 and R_2 since firm A is only concerned with the probabilities of refinancing, β_1 and β_2.)

We now establish that if firm B enters, in equilibrium $(\beta_1^*, \beta_2^*) = (0,1)$ and firm A preys. Given any strategy by firm A (and hence any probability of π_1, μ or θ) it is optimal to set $(\beta_1, \beta_2) = (0,1)$, $R_1 = \pi_1$, and $R_2 = \bar{\pi}$; this is a dominant strategy. Firm A's optimal response is then to prey. This forms the unique Nash equilibrium if firm B enters.

With observable contracts, firm B can credibly precommit to a contract that deters predation. But, when contracts are not observable, so that A must conjecture what contract B signed, firm B always wishes to set $(\beta_1, \beta_2) = (0,1)$; no precommitment is possible.

Thus, if contracts are not observable, firm B will enter provided its profits upon entry, $\pi_1 - F + (1-\mu)(\bar{\pi} - F)$, are positive. Note that firm B enters (weakly) less often when contracts are unobservable. These results are summarized in the following proposition.

PROPOSITION 3: *Firm B enters if and only if* $\pi_1 - F + (1-\mu)(\bar{\pi} - F) > 0$. *If the firm enters, the contract is as given in Proposition* 1 *and firm A preys.*

This result, like Proposition 2, is sensitive to the assumption that $\pi_1 < F$. If instead $\pi_1 > F$, in general the optimal (β_1, β_2) pair is a function of the probability of π_1 and hence whether firm A preys. Therefore, $(\beta_1, \beta_2) = (0,1)$ is not necessarily a dominant strategy. Given that firm A preys and the probability of π_1 is μ, the optimal response by firm B may be to set $\beta_1 = \beta_2 = 1$. But if this is true, it is not in the interest of firm A to prey. Similarly, if firm A does not prey, it may be optimal to set $(\beta_1, \beta_2) = (0,1)$; but then firm A has an incentive to prey. As a result, there may be no pure strategy equilibrium. One can show, however, that there is a mixed strategy equilibrium in which firm A preys with positive probability and firm B sets $\beta_1 = 1 - \Delta$ and $\beta_2 = 1$.[22] Thus, in equilibrium, the investor partially deters predation.

III. Concluding Remarks

The central argument of this paper is that agency problems in financial contracting can give rise to rational predation. The financial contract that minimizes agency problems also maximizes rivals' incentives to prey. As a result, there is a tradeoff between deterring predation and mitigating incentive problems: reducing the sensitivity of the refinancing decision to the firm's performance discourages predation, but exacerbates the incentive problem. In equilibrium, whether financial contracts deter predation depends on the relative importance of these two effects.

Our theory of predation departs from the existing literature which views predation as an attempt to convince rivals that it would be unprofitable to remain in the industry. In our model, everyone knows it is profitable for the rival to remain in the industry. Nevertheless, predation induces liquidation and exit because it adversely affects the agency relationship between the rival's investors and manager.

Although our model narrowly focuses on predation, we believe that the model provides a useful starting point to analyze a broader set of issues concerning competitive interaction among firms with agency problems and financial constraints. For example, our model suggests that an important determinant of product-market success is the degree to which firms can finance investment with internally generated funds. This is in

---

will have an incentive to privately annul the advertised contract and write a new one.

[22] More details of this argument can be found in an earlier version of our paper issued as a Sloan Working Paper No. 1986-88.

contrast to standard models of dynamic competition in which the only relevant consideration is the total capital stock and not the way in which it was acquired. The implictions of our model are consistent with Gordon Donaldson's (1984) findings that one reason managers prefer internal sources of funds is that it enhances their ability to compete in product markets.

In addition, our model suggests that certain types of product-market competition can increase managerial incentive problems within the firm. As implied by our model, the reliance on external financing exposes the firm to cutthroat competition. This may force the firm to rely more on internal sources of capital than on external ones. But, this reduces the extent to which outside investors monitor the firm and increases the possibility of managerial slack. Thus, external financing comes with costs and benefits: on the one hand, it disciplines management, but on the other, it makes the firm vulnerable in its product markets.

APPENDIX

LEMMA 1: *The incentive-compatibility constraint* (1) *is binding.*

PROOF: Suppose to the contrary that (1) is slack and that the only constraints are (2) and (2′). We establish that the optimal solution to this relaxed program violates (1).

First note that since (1) is slack, we need not be concerned with the effect of $\{\beta_1, R_1, R_1\}$ on the optimal choice of $\{\beta_2, R_2, R_2\}$ and vice versa. Thus, the maximization problem can be written:

$$\underset{\{\beta_i, R_i, R^i\}}{\text{Maximize}}\, R_i + \beta_i R^i - \beta_i F$$

subject to

$$R_i \le \pi_i \tag{A1}$$

$$R_i + R^i \le \pi_i + \pi_1 . \tag{A2}$$ □

At an optimum to this program, $R_i = \pi_i$ and $R^i = \pi_1$. This is true in the case where $\beta_i < 1$ because given the constraint on the total payments (A2), it is optimal to shift more of the payment to the first period when it will be received with certainty. If $\beta_i = 1$, any division of payments satisfying $R_i + R^i = \pi_i + \pi_1$ is optimal and we may as well set $R_i = \pi_i$ and $R_i = \pi_1$. (Note that given $\beta_i = 1$ the division of payments between R_i and R^i has no effect on the incentive-compatibility constraint.)

The incentive constraint (1) therefore simplifies to

$$\beta_2(\bar{\pi} - \pi_1) \ge \pi_2 - \pi_1 + \beta_1(\bar{\pi} - \pi_1).$$

It is easily seen that for all feasible values of β_1 and β_2 the inequality cannot be satisfied. Thus the incentive constraint is violated at the optimum of the relaxed program, establishing the contradiction.

LEMMA 2: $R^1 = R^2 = \pi_1$ *is a part of an optimal contract.*

PROOF: Substituting the incentive constraint (1) into the objective function yields the new objective function:

$$-F + R_1 + \beta_1[R^1 - \theta F - (1-\theta)\bar{\pi}]$$
$$+ (1-\theta)\beta_2(\bar{\pi} - F).$$

It follows that $\beta_2 = 1$. Hence only the sum, $R^2 + R_2$, and not the individual values R^2 and R_2 affects the objective function and the incentive constraint. Thus we can set $R^2 = \pi_1$. If $\beta_1 = 1$ the same can be said for R_1 and R^1. If $\beta_1 < 1$, R_1 and R^1 will be chosen to maximize $R_1 + \beta_1 R^1$ since it simultaneously maximizes the objective function and relaxes the incentive constraint. This expression is maximized subject to the limited-liability constraints by setting $R_1 = R^1 = \pi_1$. This completes the proof. □

REFERENCES

Benoit, Jean-Pierre, "Financially Constrained Entry in a Game with Incomplete Information," *Rand Journal of Economics*, Winter 1984, *15*, 490–99.

Bolton, Patrick and Scharfstein, David S., "Agency Problems, Financial Contracting, and Predation," Sloan Working Paper 1988-86, MIT, February 1988.

Brander, James A. and Lewis, Tracy R., "Oligopoly and Financial Structure," *American Economic Review*, December 1986, *76*, 956–70.

Donaldson, Gordon, *Managing Corporate Wealth*, New York: Praeger, 1984.

Fershtman, Chaim and Judd, Kenneth L., "Equilibrium Incentives in Oligopoly," *American Economic Review*, December 1987, *77*, 927–40.

______, "Strategic Incentive Manipulation in Rivalrous Agency," Northwestern University, 1987.

Flannery, Mark, "Asymmetric Information and Risky Debt Maturity Choice," *Journal of Finance*, March 1986, *41*, 19–38.

Fudenberg, Drew and Tirole, Jean, "A 'Signal-Jamming' Theory of Predation," *Rand Journal of Economics*, Autumn 1986, *17*, 366–76.

Gale, Douglas and Hellwig, Martin, "Incentive-Compatible Debt Contracts: The One Period Case," *Review of Economic Studies*, October 1985, *52*, 647–64.

Gertner, Robert, Gibbons, Robert and Scharfstein, David, "Simultaneous Signalling to the Capital and Product Markets," *Rand Journal of Economics*, Spring 1988, *19*, 173–90.

Hermalin, Benjamin, "Adverse Selection and Contract Length," MIT, 1986.

Jensen, Michael C., "Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers," *American Economic Review*, May 1986, *76*, 323–29.

Joskow, Paul L. and Klevorick, Alvin K., "A Framework for Analyzing Predatory Pricing Policy," *Yale Law Journal*, December 1979, *89*, 213–70.

Katz, Michael L., "Game-Playing Agents: Contracts as Precommitments," Princeton University, 1987.

Maksimovic, Vojislav, "Capital Structure in a Stochastic Oligopoly," unpublished doctoral dissertation, Harvard University, 1986.

McGee, John, "Predatory Price Cutting: The Standard Oil (N.J.) Case," *Journal of Law and Economics*, October 1958, *1*, 137–69.

Milgrom, Paul and Roberts, John, "Limit Pricing and Entry under Incomplete Information," *Econometrica*, March 1982, *50*, 443–60.

Myerson, Roger B., "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics*, June 1982, *10*, 67–81.

Ordover, Janusz A. and Willig, Robert D., "An Economic Definition of Predation: Pricing and Product Innovation," *Yale Law Journal*, November 1981, *91*, 8–53.

Sahlman, William, "Note on Financial Contracting: Deals," Harvard Business School, 1986.

Saloner, Garth, "Predation, Mergers, and Imperfect Information," *Rand Journal of Economics*, Summer 1987, *18*, 165–87.

Salop, Steven and Shapiro, Carl, "A Guide to Test-Market Predation," Princeton University, 1982.

Sappington, David E. M., "Limited Liability Contracts between Principal and Agent," *Journal of Economic Theory*, February 1983, *29*, 1–21.

Scharfstein, David S., "A Policy to Prevent Rational Test-Market Predation," *Rand Journal of Economics*, Summer 1984, *15*, 229–43.

Telser, Lester G., "Cutthroat Competition and the Long Purse," *Journal of Law and Economics*, October 1966, *9*, 259–277.

Tirole, Jean, *The Theory of Industrial Organization*, Cambridge: MIT Press, 1988.

Townsend, Robert, "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, October 1979, *21*, 265–93.

______, "Optimal Multiperiod Contracts and the Gain from Enduring Relationships Under Private Information," *Journal of Political Economy*, December 1982, *90*, 1166–86.

Vickers, John, "Delegation and the Theory of the Firm," *Economic Journal*, Supplement 1985, *95*, 138–47.

Horizontal Mergers: An Equilibrium Analysis

By JOSEPH FARRELL AND CARL SHAPIRO*

We analyze horizontal mergers in Cournot oligopoly. We find general conditions under which such mergers raise price, and show that any merger not creating synergies raises price. We develop a procedure for analyzing the effect of a merger on rivals and consumers and thus provide sufficient conditions for profitable mergers to raise welfare. We show that traditional merger analysis can be misleading in its use of the Herfindahl Index. Our analysis stresses the output responses of large firms not participating in the merger. (JEL 022,612)

Mergers between large firms in the same industry have long been a public policy concern. In the United States, Section 7 of the Clayton Act (as amended by the Celler-Kefauver Act) prohibits mergers that "substantially decrease... competition or tend... to create a monopoly." Under the Hart-Scott-Rodino Act, large firms must report any proposed substantial merger to the Department of Justice and the Federal Trade Commission, which evaluate the merger's likely effect on competition and can choose to permit or to oppose it.

In evaluating proposed mergers, federal antitrust officials generally apply rules summarized in the Department of Justice's Merger Guidelines (1984).[1] An important part of merger analysis under these guidelines involves estimating the effect of a proposed merger on market concentration. In particular, the analyst is instructed to pay careful attention to the initial *level* of concentration in the industry and the predicted *change* in concentration due to the merger. Roughly speaking, the guidelines permit mergers that will not increase concentration by very much or that will leave it low even after the merger. This reflects a view that anticompetitive harm is an increasing function of concentration, which is measured using the Herfindahl-Hirschman index, H, defined as the sum of the squares of the firms' market shares.

The Merger Guidelines, while surely more sophisticated than what they replaced, are not based on explicit analysis of how a merger will affect equilibrium output and welfare. This theoretical shortcoming leads to two basic problems with the guidelines' use of concentration measures.

The first such problem is the curious rule that the guidelines use to estimate the effect of a merger on H. This rule takes the initial market shares of the merging firms, s_1 and s_2, and assumes that the new entity's market share will be $s_1 + s_2$, so that the merger will increase H by $(s_1+s_2)^2-(s_1^2+s_2^2)=2s_1s_2$. But if indeed all firms maintain their premerger outputs, then the merger will affect neither consumers nor nonparticipant firms, so it will be socially desirable if and only if it is privately profitable. If, as is more likely, outputs change in response to the merger, then the $2s_1s_2$ formula is wrong. Equilibrium analysis, such as we provide below, is neces-

---

*University of California at Berkeley and Princeton University. Some of the results in this paper were in Farrell and Shapiro (1988), which has been split into the current paper and Farrell and Shapiro (1990). We thank Peter Bronsteen, Avinash Dixit, Gene Grossman, Michael Katz, Louis Kaplow, F. William McElroy, Rob Porter, Stephen Salant, Marius Schwartz, John Vickers, Robert Willig, and seminar participants at the Antitrust Division, Department of Justice, the Federal Trade Commission, U.C. at Berkeley, U.C. at San Diego, Maryland, Northwestern, Princeton, Columbia, Harvard, and Stanford for their comments. We especially thank Steve Salop for valuable suggestions. Both authors acknowledge the financial support of the National Science Foundation, Farrell under grant IRI 87-12238, and Shapiro under grant no. SES-8606336. Farrell also thanks the Hoover Institution and Shapiro the John M. Olin Foundation.

[1] For a discussion of these guidelines, see the symposium in the *Journal of Economic Perspectives*, Fall 1987.

sary to compute the change in H in a logically consistent manner.

The second problem is deeper and more serious. Implicitly, the guidelines assume a reliable (inverse) relationship between market concentration and market performance. In particular, the entire approach presumes that a *structural* change, such as a merger, that increases the equilibrium value of H also systematically reduces equilibrium welfare, W, defined as the sum of consumer and producer surplus or equivalently the difference between gross consumer benefits and production costs.

Is there in fact such a reliable relationship between changes in market concentration and changes in economic welfare? In some very special circumstances, there is. For example, if n equally efficient firms with constant unit cost compete as Cournot oligopolists, then W is increasing in n and $H = 1/n$, so there is a rigid (inverse) relationship between H and W as n varies.[2] But if the competing firms are not equally efficient, or if there are economies of scale, there is no reason to expect that concentration and welfare will move in opposite directions in response to a merger.[3]

The dangers of identifying changes in H with changes in W are suggested by the fact that, starting at a Cournot equilibrium, welfare rises with a small change in firms' outputs if and only if

$$\frac{dX}{X}+\frac{1}{2}\frac{dH}{H}>0, \tag{1}$$

where X is aggregate output. (See the Appendix for a derivation of this condition.) Of course, increases in output ($dX > 0$) tend to raise welfare, since price exceeds marginal cost in equilibrium. But if a firm with a large market share increases its output, then H, X, and W will all rise. And condition (1) shows that for any given (percentage) change in X, welfare is more likely to rise if H *increases*; it is even possible to have $dX < 0$, $dH > 0$, and yet $dW > 0$.[4]

How can increases in the concentration be associated with increases in welfare? In Cournot equilibrium, larger firms have lower marginal costs, so welfare is enhanced if a fixed total output X is shifted toward them and away from smaller, less efficient firms. But such shifts will increase concentration.

This observation is not a theoretical curiosity. Critics of U.S. antitrust policy have long argued that large firms may be large because they are efficient. If so, then economic welfare may be enhanced if these efficient firms acquire more of the industry's productive capital and thus increase their market share.[5] One means to do this is by buying the assets of smaller, less efficient rivals. Any useful theory of oligopoly and horizontal mergers should account for the role that mergers may play in this process.

Given the complex relationship between concentration, output, and welfare, a careful analysis of the welfare effects of mergers is badly needed. The few existing theoretical analyses of the effects of horizontal mergers have used very special models, and their insights of course reemerge below.[6] But our goal is more ambitious: we use Cournot oligopoly theory, with quite general cost and demand functions, to study the output and welfare effects of mergers. At a theoretical level, we develop some techniques for analyzing welfare changes in Cournot markets;

---

[2] In this model, aggregate output, X, is also a sufficient statistic for welfare, as it is more generally if total production costs depend only on aggregate output, and not on the distribution of output across firms.

[3] For further discussion of the relationship between changes in H and changes in W, see Farrell and Shapiro (1990).

[4] For example, consider a Cournot duopoly in which firm 1 has low marginal cost and hence a large market share, and firm 2 has high marginal cost and is small. Closing down firm 2 will scarcely reduce welfare (since its marginal cost is close to the price). And if firm 1 expands at the same time, the net welfare effect will be favorable, since firm 1's marginal cost is distinctly less than the price. Such an output shift would lower X, raise H and yet raise W.

[5] For an influential statement of this view, see for instance Harold Demsetz (1973, 1974).

[6] Steven Salant, Sheldon Switzer, and Robert Reynolds (1983), Raymond Deneckere and Carl Davidson (1985), and Martin Perry and Robert Porter (1985) are the most prominent papers.

these techniques apply not only to mergers but more generally. At a policy level, we find surprisingly general sufficient conditions for when a merger should be approved or prohibited.

Our policy analysis is in two parts. First, we study a merger's effect on price. This inquiry will be most telling for those antitrust practitioners who see consumer welfare as the sole objective of antitrust policy. We provide a necessary and sufficient condition for a merger to raise price (Proposition 1) and show in general that mergers in Cournot oligopoly raise price if they generate no synergies between the merging firms (Propositions 2 and 3). We also show that firms with large market shares must achieve impressive synergies or scale economies if their merger is to reduce price.

Second, we study a merger's effect on welfare, W. Here we emphasize the *external* effects on consumers and on nonparticipant firms. In this, our work is squarely in the tradition of mainstream economic analysis, whose presumption has always been that a debate about intervention should focus on externalities.[7] Write $\Delta\pi^I$ for the change in joint profits of the merging firms (the "insiders"), and ΔW for the change in total welfare associated with a merger. We analyze the next external effect of the merger on rival firms ("outsiders") and on consumers, $\Delta W - \Delta\pi^I \equiv \Delta\pi^O + \Delta CS$. Since any proposed merger is presumably privately profitable, it will also raise welfare if it has a positive external effect—a condition that is, perhaps surprisingly, quite often satisfied. We thus find sufficient conditions for profitable mergers to raise welfare (Proposition 5).

Our emphasis on the external effect also has a great practical advantage. To assess the externality turns out to require much less information than to assess the overall welfare effect, since the effect on insiders' profits depends on internal cost savings. Such cost savings are often hard to observe, and as a result are given little weight in the Merger Guidelines and in many practitioners' thinking, even though, as Oliver Williamson (1968) stressed, they may well be very important in a merger's overall welfare effect. Thus, Lawrence White (1987, p. 18) wrote:

> Efficiencies are easy to promise, yet may be difficult to deliver. All merger proposals will promise theoretical savings in overhead expense, inventory costs, and so on; they will tout "synergies."

Franklin Fisher (1987, p. 36) concurred:

> The burden of proof as to cost savings or other offsetting efficiencies, however, should rest squarely on the proponents of a merger, and here I would require a very high standard [of proof]. Such claims are easily made and, I think, often too easily believed.

This emphasis on the externality is our first innovation in merger policy analysis, and it also leads us into our second. Although the externality is easier to work with than the total welfare effect, directly signing the externality is mathematically difficult even in the simplest cases, and completely intractable in general. To overcome this, we introduce differential techniques. The external effect of a merger involving a change ΔX_I in the merging firms' joint output can be calculated as the integrated external effect of *small* changes dX_I. We call such a small change an "infinitesimal merger." It turns out to be easy to sign the external effect of an infinitesimal merger (Proposition 4), and in fairly general circumstances we can sign the external effect of a merger (Proposition 5).

Our paper is organized as follows. In Section I, we recall some results from standard Cournot oligopoly theory. In Section II, we analyze the output and price effects of a merger among Cournot oligopolists, and find broad sufficient conditions for a merger to raise the equilibrium price. In Section III, we calculate the external effect of an infinitesimal merger, and show that it depends only on the merging firms' joint market share and on the output responsiveness and market shares of nonparticipant firms. We then give conditions under which privately profitable mergers raise total welfare. Finally, in Section IV, we use our methods to provide welfare analyses of two other oligopoly problems: commitment in Cournot competition, and the optimal level of an import quota. A conclusion follows.

---

[7] Of course, mergers typically do not generate externalities in the usual sense of the term. With imperfectly competitive markets, however, a change in the behavior of merging firms does affect the welfare of consumers and other firms. Throughout the paper we take "externality" to mean $\Delta W - \Delta\pi^I$.

I. Cournot Oligopoly

We use the traditional model of Cournot oligopoly with homogeneous goods. Demand is given by $p(X)$, where p is price, X is industry output, and $p'(X) < 0$; we write ε for the absolute value of the elasticity of demand, $\varepsilon(X) \equiv -p(X)/Xp'(X)$. The number of firms, n, is exogenous (although of course it changes with a merger), reflecting some important barriers to entry.[8] We denote firm i's cost function by $c^i(x_i)$, where x_i is firm i's output. For notational ease, we write $c^i \equiv c^i(x_i)$ for firm i's total cost and $c_x^i \equiv c_x^i(x_i)$ for firm i's marginal cost. Importantly, we permit the firms to differ in efficiency.

In the Cournot equilibrium, each firm i picks its output x_i to maximize its profits, given its rivals' outputs. Writing $y_i \equiv \sum_{j \neq i} x_j = X - x_i$ for aggregate output of all firms other than firm i, firm i's profits are $\pi^i(x_i, y_i) \equiv p(x_i + y_i)x_i - c^i(x_i)$. Firm i's first-order condition, $\partial \pi^i / \partial x_i = 0$, is

$$(2) \quad p(X) + x_i p'(X) - c_x^i(x_i) = 0, \qquad i = 1, \ldots, n.$$

A Cournot equilibrium is a vector $(x_1, \ldots, x_n)$ such that equation (2) holds for all n firms. We denote firm i's market share by $s_i \equiv x_i / X$.

Comparing two firms i and j, the Cournot equilibrium conditions, (2), tell us that $x_i > x_j$ if and only if $c_x^i < c_x^j$. In equilibrium, larger firms have lower marginal costs. In any Cournot equilibrium in which different firms produce different quantities, marginal costs differ across firms, so that costs are not minimized given the aggregate output level, and consequently aggregate output, X, is *not* in general a sufficient statistic for welfare.

Throughout the paper, we make two weak assumptions on the Cournot equilibrium; we require both to hold throughout a relevant range, as will become clear below. First, we assume that each firm's reaction curve slopes downward. Equivalently, an increase in rivals' output, y_i, lowers firm i's marginal revenue:

$$(3) \quad p'(X) + x_i p''(X) < 0, \qquad i = 1, \ldots, n.$$

Inequality (3) is a very weak assumption and is standard in Cournot analysis; see Avinash Dixit (1986) and Shapiro (1989). It holds if the industry demand curve satisfies $p'(X) + Xp''(X) < 0$.[9]

Second, we assume that each firm's residual demand curve, $p(\cdot + y_i)$, intersects its marginal cost curve from above. This is equivalent to

$$(4) \quad c_{xx}^i(x_i) > p'(X), \qquad i = 1, \ldots, n.$$

Condition (4) is surely met if marginal cost is nondecreasing, i.e., if $c_{xx}^i \geq 0$. It is among the weaker known stability conditions for Cournot equilibrium (Dixit, 1986).

We now note some comparative-statics properties of Cournot equilibria that will be important below. Consider the effect of a change in rivals' aggregate output, y_i, on firm i's output. From equation (2), the slope of firm i's reaction schedule is given by

$$\frac{dx_i}{dy_i} \equiv R_i = -\frac{p' + x_i p''}{2p' + x_i p'' - c_{xx}^i}.$$

---

[8] Our analysis can easily accommodate entry by, or the existence of, price-taking fringe firms, if we reinterpret the demand curve $p(X)$ as the residual demand curve facing the oligopolists that we model. What we are ruling out is entry by additional large firms that behave oligopolistically.

[9] If $p'' \leq 0$, then (3) surely holds. If $p'' > 0$, then $p' + Xp'' > p' + x_i p''$.

From condition (3) and firm i's second-order condition, $R_i < 0$; with (4) we also have $-1 < R_i < 0$. That is, if its rivals jointly expand production, firm i contracts, but by less than its rivals' expansion. From $dx_i = R_i dy_i$ we have $dx_i(1+R_i) = R_i(dx_i + dy_i) = R_i dX$, or

$$dx_i = -\lambda_i dX, \tag{5}$$

where

$$\lambda_i \equiv -\frac{R_i}{1+R_i} = -\frac{p'(X)+x_i p''(X)}{c^i_{xx}(x_i)-p'(X)}.$$

Below, it will prove easier to work with λ_i instead of R_i. Under conditions (3) and (4), $\lambda_i > 0$. The λ_i's are important below, so we pause here to show how they can be expressed in terms of more familiar elasticities. Write $E \equiv -Xp''(X)/p'(X)$ for the elasticity of the slope of the inverse demand curve, and $\mu_i \equiv x_i c^i_{xx}/c^i_x$ for the elasticity of firm i's marginal cost, c^i_x, with respect to its output, x_i.[10] Then

$$\lambda_i = \frac{s_i - s_i^2 E}{s_i + \mu_i(\varepsilon - s_i)}. \tag{6}$$

With constant elasticity of demand ε, $E = 1+1/\varepsilon$, so λ_i can be expressed solely in terms of s_i, μ_i, and ε. With linear demand and constant marginal costs, $E = 0 = \mu_i$, so $\lambda_i = 1$.

Finally, we prove in the Appendix and record here a fact[11] about the response of all other firms to an output change by one firm:

LEMMA: *Consider an exogenous change in firm* 1*'s output, and let the other firms' outputs adjust to re-establish a Cournot equilibrium among themselves. If firms' reaction curves slope downward* (*condition* (3)), *and if the stability condition* (4) *holds, then aggregate output moves in the same direction as firm* 1*'s output, but by less.*

Note that for the Lemma to hold it is *not* necessary that firm 1 behave as a Cournot oligopolist.

II. Price Effects of Horizontal Mergers

We model a merger as a complete combination of the assets and of the control of the merging firms, whom we call the "insiders." After the merger, a new Cournot equilibrium is established between the merged entity M and the nonparticipant firms, whom we call the "outsiders." In this section, we examine the effect of a merger on aggregate output, X. This is, of course, the central question if merger analysis is concerned only with consumer welfare (ignoring consumers' ownership of profits). As we shall see in Section III, it is also an essential component of an analysis of overall economic welfare, W.

Mergers differ enormously in the extent to which productive assets can usefully be recombined, and in the extent to which output decisions can usefully, or anticompetitively, be coordinated. At one extreme, consider a production technology in which all firms have constant and equal marginal costs, and the merged entity has the same costs. In this special case, mergers are purely anticompetitive: there is no other motive. For a slightly rosier view, recall that by equation (2), firms' marginal costs typically differ in Cournot equilibrium, so that a merger may offer an opportunity to *rationalize production*—that is, without changing total output, to shift output to the facility with lower marginal cost. A still sunnier view is that mergers may create synergies. For example, two firms that own complementary patents may combine and produce much more efficiently than either could alone (without a licensing agreement).

For a theory of horizontal mergers to be useful for policy purposes, it should be general enough to allow for all these possibilities, which can be captured in assumptions about the relationship between the merged entity's cost function, $c^M(\cdot)$, and those of the insiders. Our theory is very general in this regard, as we make no a priori assumptions on $c^M(\cdot)$ beyond those implied by conditions (3) and (4). Our first proposition gives a necessary and sufficient condition on $c_x^M(\cdot)$ for equilibrium output to fall with the merger.

---

[10] Note that μ_i is the inverse of the elasticity of firm i's competitive supply curve.

[11] Dixit (1986) notes this property of Cournot equilibrium.

PROPOSITION 1: *A merger of a group of firms in Cournot oligopoly raises price if and only if M's markup would be less than the sum of the pre-merger markups at its constituent firms, were M produce just as much as its constituent firms together did before the merger.*

PROOF:

See the Appendix. □

In the typical situation where two firms (1 and 2) are merging, price will fall if and only if $p - c_x^M > (p - c_x^1) + (p - c_x^2)$, where p is the pre-merger price, c_x^1 and c_x^2 are measured at pre-merger output levels, $\bar{x}_1$ and $\bar{x}_2$, and c_x^M is measured at output $\bar{X}_M \equiv \bar{x}_1 + \bar{x}_2$. Equivalently, price will fall if and only if

$$c_x^2 - c_x^M > p - c_x^1. \tag{7}$$

By condition (7), M must enjoy *substantially* lower marginal costs than did its constituent firms, if price is to fall. Furthermore, the required reduction in marginal costs is larger, the larger were the pre-merger markups of the merging firms; and those markups were, by (2), proportional to pre-merger market shares.

Using equation (2), we can express (7) in terms of pre-merger variables that are relatively easy to observe:

$$c_x^M < p\left(1 - \frac{s_1 + s_2}{\varepsilon}\right).$$

This version shows how much less than current prices the merged firm's marginal costs must be, if price is to fall. We now present some illustrative calculations to draw out some of the implications of condition (7) and Proposition 1.

A. *Mergers with No Synergies*

In many mergers, the insiders (firms $i \in I$) cannot recombine assets to improve their *joint* production *capabilities*. After the merger, the combined entity M can perhaps better allocate outputs across facilities ("rationalization"), but M's production *possibilities* are no different from those of the insiders (jointly) before the merger. In this case we say that the merger "generates no synergies." Formally, with no synergies,

$$c^M(x) \equiv \min\left\{ \sum_{i \in I} c^i(x_i') \mid \sum_{i \in I} x_i' = x \right\}. \tag{8}$$

This holds, for instance, in the constant average cost model of Steven Salant, Sheldon Switzer, and Robert Reynolds (1983) and in the quadratic cost model of Martin Perry and Robert Porter (1985).

PROPOSITION 2: *If a merger generates no synergies, then it causes price to rise.*

PROOF:

See the Appendix. □

When will our no-synergies condition, (8), apply? For illustration, write firm i's cost function as $c^i(x_i) = \theta_i \phi(x_i, k_i)$, where $\phi(\cdot,\cdot)$ is a short-run variable cost function, k_i measures the amount of a possibly fungible capital good employed at firm i, and θ_i inversely measures "knowledge" at firm i.

With this form of the cost function we can distinguish three types of cost savings from a merger. (1) Participants may rationalize output across their facilities (recall that pre-merger marginal costs are typically unequal); this consists of changing the x_i's but not the k_i's or the θ_i's. (2) They may shift capital across their facilities, changing the distribution of the k_i's (but not the total capital stock available). (3) They may learn from each other, that is, share techniques, patents, or management skills; learning will change the θ_i's.

Proposition 2 tells us that efficiencies of the first kind, while of course desirable, cannot suffice to make output rise instead of fall.[12] Efficiencies of the second kind apply only in the "short run"—that is, when capital cannot be brought in from the outside. These efficiencies are further limited by the observation that equation (8) holds even with mobile capital, if the firms produce according to the same long-run cost function (i.e., they have the same level of knowledge) and if this common technology exhibits constant returns to scale. In that case, even if capital can be moved, there is no point in doing so: equivalent results can be achieved by reallocating output across facilities. This proves

PROPOSITION 3: *Suppose that a merger involves no learning. Then, in the long run, the merger will raise price. In the short run, it will raise price* (1) *if capital is immobile across facilities, or* (2) *if all merging firms are equally efficient and their long-run production function exhibits constant returns to scale.*

Thus, a merger can raise output and make consumers better off only if it permits the merging firms to exploit economies of scale or if the participants learn from it. In the next two subsections, we illustrate how large these effects must be for a merger to lower price.

B. *Mergers with Economies of Scale*

Consider the possibility that some form of "capital" may be best recombined after merger. In particular, suppose that it pays—and is possible—to bring together all the new entity's capital rather than leaving it divided among its plants (formerly, firms) in its pre-merger configuration. This will always be desirable if there are economies of scale; whether capital is mobile is of course a technical question. How much economies of scale are needed for a merger to increase output and reduce price?

Consider for illustration a merger between two *ex ante* identical firms each with a pre-merger market share of s. There is a common variable cost function $c(x,k)$, and each firm owns an amount k of capital prior to the merger. The merger permits the firms to combine their capital as well as reallocate outputs; and with economies of scale, M will do so. The one combined facility will then produce with the variable cost function $c(\cdot,2k)$.

We show in the Appendix that the merger will lower price if and only if

$$(9)\quad c_x(2x,2k) \le \left[1-\frac{s}{\varepsilon - s}\right] c_x(x,k).$$

To illustrate (again, see the Appendix for details), let $c(x,k)$ be dual to the production function $f(L,k)=k^aL^a$. If $s=0.2$ and $\varepsilon=1$, then (9) requires $a \ge 1/\log_2 3$: twice the inputs must produce at least $2^{2/\log_2 3} \approx 2.4$ times the output.

C. *Mergers with Learning*

A merger may enhance efficiency at some or all of the merging facilities: one facility may learn from its partner's patents, management expertise, etc. How much such learning is needed for a merger to increase output and reduce price?

Suppose that firms 1 and 2 merge. Suppose for simplicity that capital cannot be reallocated across firms,[13] so we can write firm i's cost function as $\theta_i\varphi(x_i)$; suppose further that marginal costs are nondecreasing. In the Appendix we show that, for price to fall, the merger must either reduce θ_1 by at least a factor $s_2/(\varepsilon - s_1)$ or reduce θ_2 by at least a factor $s_1/(\varepsilon - s_2)$. In the symmetric case with $s_1 = s_2 = 0.2$ and $\varepsilon = 1$, this means that at least one plant must achieve a reduction in θ of at least 25 percent as a result of merger if price is to fall.

D. *Policy Implications*

Propositions 1 to 3, and our illustrative calculations in subsections B and C, support the presumption that an oligopolistic merger will reduce aggregate industry output, and point to the nature and degree of synergies or scale economies that are required to overturn this presumption. If the merger changes the behavioral mode in the industry—from Cournot behavior to something more collusive—then the presumption is even stronger that price will rise.

---

[12] After our paper was revised, we became aware of the work of F. William McElroy (1988), who proves a result similar to our Proposition 2.

[13] Equivalently, consider the long run, in which case capital does not appear in the cost function at all.

We have identified the factors that determine a merger's effect on price in Cournot equilibrium. In particular, the larger are the market shares of the participating firms, or the smaller is the industry elasticity of demand, the greater must be the learning effects or scale economies in order for price to fall. It is perhaps encouraging that these are exactly the factors that the Merger Guidelines instruct antitrust officials to consider.

Some antitrust scholars and practitioners, including the National Association of Attorneys General, argue that the objective of antitrust policy is to maximize consumers' surplus.[14] Under this view, mergers should be blocked if and only if they are expected to raise price. Propositions 1, 2, 3, and our illustrative calculations point to the relevant factors in making such an assessment.[15] We find that rather impressive synergies—learning, or economies of scale—are typically necessary for a merger to reduce price.

But few economists would accept that antitrust policy should aim to maximize consumer surplus alone, with no weight given to profits. Although direct stock ownership is concentrated among the relatively rich, many people are indirect shareholders, for instance through pension funds, so it seems inappropriate to ignore profits entirely. In a conventional economic view, the proper goal of antitrust policy is to maximize overall market efficiency or welfare, and in this case further analysis is required.

III. Welfare Effects of Horizontal Mergers

We will now analyze the welfare effects of horizontal mergers. Unfortunately, it is hard to compare pre- and post-merger allocations directly, even in the simplest special cases. But we will show that a merger's effect on the welfare of nonparticipating firms and consumers, although generally unavailable in closed form, can be expressed as the integral of a relatively simple integrand, namely the external effect of what we will call an "infinitesimal merger." We obtain some powerful results by examining that integrand.

A merger generally changes all firms' outputs in equilibrium. But consumers care only about the net effect on aggregate output, ΔX, and (in Cournot equilibrium) rivals care only about the change in equilibrium output by the merging ("insider") firms, ΔX_I, not about what caused that change. Therefore, in examining the external effects of a merger, once we know the equilibrium change ΔX_I we need no information about what went into that change: we can simply ask how outsider firms respond and what is the effect on their profits and on consumer surplus, and in doing so we can treat ΔX_I as exogenous.

Moreover, the effect of ΔX_I can be decomposed into the integral of the effects of infinitesimal changes dX_I that make up ΔX_I. We call such a small change in insiders' output, dX_I, an "infinitesimal merger" if it has the same sign as the change in X_I consequent on the merger among the insiders. For mathematical convenience, we can think of a merger as the composite of many such infinitesimal mergers.[16]

---

[14] For example, in *Proctor & Gamble*, the Supreme Court stated that "...Possible economies cannot be used as a defense to illegality. Congress was aware that some mergers which lessen competition may also result in economies, but it struck the balance in favor of protecting competition." See the discussion in Williamson (1968), for instance.

[15] Note that under this view of antitrust policy, rival firms should not have standing to sue to block a proposed merger, since their interests are diametrically opposed to consumers'. Indeed, when rival firms contest a proposed merger, arguing (ostensibly with commendable unselfishness) that the proposed merger would reduce output and should be forbidden, we should perhaps infer that they believe the opposite, and that the merger would probably benefit consumers! Sometimes rivals adopt a theory of "incipient predation," arguing that the merged entity will increase output, but only as a form of predation against them. This too is unconvincing since predation can be directly, and profitably, fought under the treble-damages provisions of the antitrust laws.

A. *External Welfare Effects of Mergers*

Consider an infinitesimal merger, dX_I. Because the marginal gross benefits of output are measured by the market price p, we have

$$dW = p\,dX_I - dc^I + \sum_{i\in O}\left[p - c_x^i\right]dx_i, \tag{10}$$

where c^I is the insiders' total cost and O is the set of outsider firms. These outsiders' output responses are given by (5), $dx_i = -\lambda_i dX$, and their markups are given by (2), $p - c_x^i = -x_i p'(X)$. Adding and subtracting $X_I dp$, and making these substitutions, we can rewrite (10) as

$$\begin{aligned} dW = {} & \left(p\,dX_I + X_I\,dp - dc^I\right) \\ & - X_I p'(X)\,dX \\ & + \sum_{i\in O} p'(X)\lambda_i x_i\,dX. \end{aligned} \tag{11}$$

In (11), the first three terms constitute the change in the insiders' joint profits, $d\pi^I$. Clearly, $d\pi^I$ involves the cost term dc^I, which is very hard for antitrust enforcement officials to observe, and any attempt to put numbers or signs to equation (11) is consequently difficult. But this troublesome term drops out when we examine the *external* welfare effect. From (11),

$$\begin{aligned} dW - d\pi^I = {} & -X_I p'(X)\,dX \\ & + \sum_{i\in O} p'(X)\lambda_i x_i\,dX, \end{aligned}$$

or

$$dW - d\pi^I = \left(\sum_{i\in O}\lambda_i x_i - X_I\right) \times p'(X)\,dX. \tag{12}$$

Defining

$$\eta \equiv \sum_{i\in O}\lambda_i x_i - X_I, \tag{13}$$

the net externality from an infinitesimal merger that induces an overall output change dX is $\eta p'(X)\,dX$, which has the sign of η if $dX < 0$. Converting η into market shares, we have:

PROPOSITION 4: *Consider any change in behavior by a subset of firms, "insiders," in an oligopolistic industry. The net external effect of this change on other firms, "outsiders," who are Cournot oligopolists, and on consumers depends only on the equilibrium change in the insiders' output, X_I. A small reduction in X_I has a net positive welfare effect on outsiders and consumers if and only if $\sum_{i\in O}\lambda_i s_i > s_I$.*

Proposition 4 underlines the importance of nonparticipant firms' responses to the change in X_I.[17] If they did not respond, that is, if $\lambda_i = 0$ for $i \in O$, then every output reduction would be bad for rivals and consumers jointly: rivals would benefit, but consumers would lose by more. This often seems to be implicitly assumed in merger policy. But in fact, as Proposition 4 shows for small arbitrary changes and Proposition 5 will show for mergers, many output-reducing changes benefit rivals more than they hurt consumers.

We now use Proposition 4 to establish our central result: a sufficient condition for a privately profitable merger to be socially beneficial. Evidently, this suggests a "safe harbor" provision for merger policy: if Cournot behavior is thought to be a good description of industry behavior and if the conditions of Proposition 5 hold, then any proposed merger should be permitted.

---

[16] An infinitesimal merger may correspond to an economic event, such as the transfer of a small amount of capital from one firm to another or the purchase by one firm of a small ownership stake in another firm. In Farrell and Shapiro (1990) we explore these changes in the ownership of assets in oligopoly. But here we use infinitesimal mergers strictly as a mathematical construct.

[17] Salant, Switzer, and Reynolds (1983) emphasize this effect in their very special model, although they do not explicitly discuss the externality. We will return to their case below.

If a merger, or other discrete change, causes the insiders' equilibrium output to change from X_I^{initial} to X_I^{final}, we have

$$\text{(14)} \quad \Delta W - \Delta\pi^I = \int_{X_I^{\text{initial}}}^{X_I^{\text{final}}} \left(\frac{dW}{dX_I} - \frac{d\pi^I}{dX_I} \right) dX_I,$$

where for each X_I, the integrand is evaluated assuming a Cournot equilibrium among outsiders given X_I. For expositional simplicity, and as suggested by Propositions 1–3, we focus on output-reducing mergers, for which $\Delta X_I < 0$. With X_I falling, we rewrite (14) as

$$\text{(15)} \quad \Delta W - \Delta\pi^I = \int_{X_I^{\text{final}}}^{X_I^{\text{initial}}} - \left(\frac{dW}{dX_I} - \frac{d\pi^I}{dX_I} \right) dX_I.$$

Using (12) and (13), we can rewrite (15) as

$$\text{(16)} \quad \Delta W - \Delta\pi^I = \int_{X_I^{\text{final}}}^{X_I^{\text{initial}}} \eta(X)[-p'(X)] \frac{dX}{dX_I} dX_I.$$

This shows that the net externality is a weighted integral of η along a path from X_I^{initial} to X_I^{final}. Consequently, if we can sign η throughout such a path, then we can sign the total external effect, $\Delta W - \Delta\pi^I$.

In Proposition 5, we give conditions sufficient for η to increase as X_I falls. When those conditions hold, it follows that if $\eta \geq 0$ before a merger, and if the merger will involve a reduction in insiders' output, then the total externality $\Delta W - \Delta\pi^I$ is surely positive.[18]

PROPOSITION 5: *Consider a proposed merger among firms* $i \in I$, *and suppose that their initial (joint) market share* s_I *does not exceed* $\sum_{i \in O} \lambda_i s_i$. *Suppose further that* p'', p''', *and* c_{xx}^i *are all nonnegative and* c_{xxx}^i *is nonpositive in the relevant ranges and for all nonparticipant firms i. Then, if the merger is profitable and would raise the market price, it would also raise welfare.*

PROOF:

See the Appendix. □

We pause now to discuss the conditions of Proposition 5. The requirement that $p'' \geq 0$ is met, for example, by all constant elasticity demand curves and of course by linear demand. The condition that marginal costs not decrease, $c_{xx} \geq 0$, is also quite widely met. For moderate sized mergers (those involving modest changes in X_I), these second-order terms will dominate the calculations, so the conclusion of Proposition 5 is likely to hold even if p''' may be negative or c_{xxx}^i may be positive. For a large merger, we need conditions on third derivatives in order to sign its external effect. This is not surprising. For example, if the outsiders face capacity constraints, so $c_{xxx} > 0$, their ability to increase output will be limited and the λ_i's must fall as the merger proceeds and their output rises. Likewise, if $p''' < 0$ then outsiders' responses to decreases in X_I may diminish as X_I does. Because these third-order effects inevitably appear, and because a merger is inherently a lumpy, non-marginal change, we believe that Proposition 5 may be the strongest "clean" result available for general cost and demand functions.

B. *Implications for Antitrust Policy*

As we remarked in the Introduction, most economists believe that interventions in the economy should be based on analysis of externalities. We might therefore hope that our explicit analysis of the externalities from a merger in Cournot oligopoly will help us discuss proper policy toward mergers.

Privately unprofitable mergers will not be proposed, so proposed mergers should be permitted unless their external effects are

---

[18] Moreover, under the conditions of Proposition 5, if $\eta < 0$ *after* the merger, that is, at $X_I = X_I^{\text{final}}$, then we know that consumers and rival firms jointly are harmed by the merger.

"sufficiently" bad to outweigh their private profitability. In particular, if a proposed merger would have a *beneficial* external effect, then it should be allowed.[19] We might be able to identify such cases using equation (15) or (16). In particular, if $\eta > 0$ at pre-merger equilibrium then there is some reason to expect the net externality $\Delta W - \Delta\pi^I$ to be positive, unless we have reason to expect η to change sign over the course of the merger (regarded as a gradual sequence of fictitious infinitesimal mergers). More rigorously, Proposition 5 gives sufficient—but far from necessary—conditions for this inference to be valid. Thus, when the conditions of Proposition 5 hold, any privately profitable merger should be permitted.

But a policy that allows only mergers with positive net external effects will be too restrictive to maximize overall economic surplus. Many profitable mergers that involve a negative "wedge" ($\Delta W - \Delta\pi^I < 0$) will nevertheless increase total surplus ($\Delta W > 0$). Although we have no formal results for this case, we hope that our externality technique may nevertheless be useful in informing merger policy, as the following diagrammatic framework may illustrate.[20] In Figure 1, a merger is represented by a point whose horizontal component is the net wedge, $\Delta W - \Delta\pi^I$, that it would generate, and whose vertical component is the private profitability, $\Delta\pi^I$. Ideally, we would like to find a policy that would permit exactly those mergers "northeast" of the negatively sloped 45° line through the origin— that is, those in regions A, B, and C; those are the "socially beneficial" mergers (those with $\Delta W > 0$).

In Figure 1, we can see why a policy of allowing only mergers with positive net wedges is too restrictive. For if such a policy were implemented, then only mergers in the northeast quadrant (region B) of Figure 1 will take place (be proposed and approved). It seems worth considering, therefore, whether there is a policy that might allow more of the desirable mergers without also permitting others that are profitable but socially undesirable. Evidently, no policy that does not involve compulsion or subsidies to merger—both of which would be enormous changes from our antitrust policy—can hope to get mergers in region A implemented. We focus, therefore, on whether a policy might be found that would permit us to distinguish proposed mergers in regions B and C (profitable and socially beneficial) from those in region D (profitable but socially harmful).

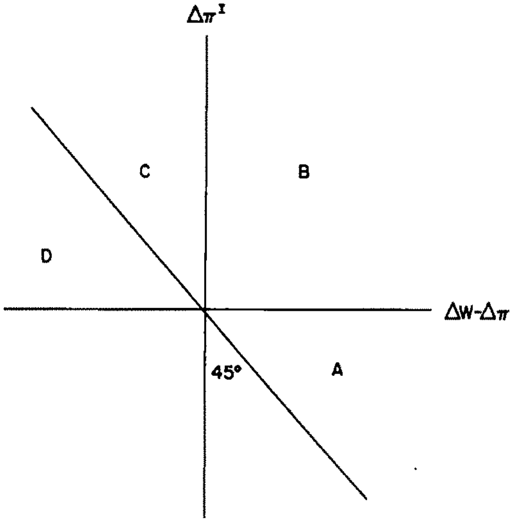


FIGURE 1

Our analysis suggests the following two-part procedure for evaluating a proposed merger. First, using (16), estimate the net wedge, $\Delta W - \Delta\pi^I$. Second, if the net wedge appears to be negative, estimate the profit effect, $\Delta\pi^I$. This might be done by observing how the aggregate stock market valuation of the participants changes with the news of the proposed merger. These estimates of the net wedge and of the profit effect can then be plotted on Figure 1, and the merger should be approved if and only if the point so plotted lies in the northeast half-plane in Figure 1.

---

[19] Note that, since nonparticipant firms' and consumers' interests concerning insiders' output are strictly opposed, a merger will never generate a Pareto improvement; we are aggregating nonparticipants' and consumers' surplus into a single measure.

[20] We are especially indebted to Steven Salop for suggesting this treatment.

We stress, however, that such a procedure would involve a number of problems, some severe. Estimating the net wedge involves estimating structural parameters such as the λ_i's, and how they will change in equilibrium with changes in X_I. The estimate of the private profitability $\Delta\pi^I$, which is needed only if the net wedge appears to be negative, is also problematic. Using the stock-market response to the announcement of the proposal assumes that no significant insider trading predates the announcement and that the stock market can accurately value the proposed change. More fundamentally, this rule would seem to create a multiple-equilibrium problem for proposed mergers that in fact lie in region C. If the market believes that the merger will be approved, then the increase in stock-market value will so impress the regulators that they will indeed approve it. If, however, the market expects regulatory opposition, then the value will rise by little, and the regulators will indeed oppose the merger!

C. *Linear Demand and Constant Costs*

In this subsection and the next, we illustrate Proposition 5 by recalling the cases analyzed by Salant et al. (1983) and by Perry and Porter (1985) and R. Preston McAfee and Michael Williams (1988).

Salant et al. assume symmetric Cournot competition among n firms, which have identical constant marginal cost c and face a linear demand curve. They consider the private and social benefits of a merger of $m+1$ of those firms—that is, a shift from Cournot to collusive behavior within the group.

For constant marginal costs and linear demand, $\lambda_i = 1$ for all i, and $\eta = X_O - X_I$. Hence, a merger among $m+1$ out of the n firms will have $\eta \geq 0$ all along the path from X_I^{initial} to X_I^{final} if the *pre-merger* outputs satisfy $X_O - X_I \geq 0$, that is, if $n - m - 1 \geq m + 1$, or $m + 1 \leq n/2$. In this model, then, any profitable merger involving at most half the industry is socially desirable.

This may seem puzzling, since, with constant unit costs, welfare declines with any merger. Salant et al. resolve this puzzle by pointing out that outsiders' responses may make the merger unprofitable, especially when m/n is small.[21] They also emphasize, however, that internalized savings (for example, fixed costs, which are perfectly consistent with constant marginal costs) may reverse that unprofitability—and hence also reverse the welfare undesirability.

Similarly, the net external effect of a merger among $m+1$ firms is surely negative if the *post-merger* outputs satisfy $X_I \geq X_O$, that is, when $1 \geq n - m - 1$, or when the number of merging firms is at least $n-1$ (merger to duopoly). This does not necessarily mean that no such merger should ever be approved, but there is certainly a case to answer.

Observe that in this model it does not matter for the externality how X_O is divided amongst the outsiders, for λ_i is independent of firm i's output or market share. In our next example we shall see that this is not generally the case.

D. *Linear Demand and Quadratic Costs*

For our second illustration, consider the quadratic cost and linear demand functions used by Perry and Porter (1985), and by McAfee and Williams (1988). Demand is given by $p(X) = A - X$, and costs are given by $c(x,k) = \frac{1}{2}x^2/k$ at all firms.[22] This variable cost function, which is dual to the Cobb-Douglas production function $x = \sqrt{Lk}$, exhibits constant returns to scale (it is homogeneous of degree one in capital and output). Hence, by Proposition 3, every merger reduces output. Proposition 5 applies, and simple calculations show that $\lambda_i = x_i/p = s_i/\varepsilon$. Consequently, a merger surely has a positive externality if

$$s_I < \frac{1}{\varepsilon} \sum_{i \in O} s_i^2, \tag{17}$$

---

[21] Gérard Gaudet and Salant (1988) study how large a group of firms must be for coordinated output reductions by the group to be profitable, despite the increased output by rivals that such reductions will induce.

[22] The analysis would be identical in a model with $p = A - bX$ and $c(x,k) = cx + ex^2/2k$, but the parameters b, c, and e would clutter our formulas.

where these are pre-merger market shares. Thus a merger is more likely to help rivals and consumers jointly if s_I is small—this conforms with standard merger analysis—and if the rest of the industry is *more* concentrated!

The reason for this surprising conclusion is that, in this model, λ_i is larger for larger firms $i\,(\lambda_i = x_i/p)$. In economic terms, if there are large outsiders, then any reduction in X_I will induce an especially large output response by them—which is just where increases in output are socially most valuable. Concentration among outsiders therefore makes their output response more socially beneficial. If λ_i were smaller at larger firms, as can easily happen,[23] then the outsiders' output response might either be weakened or less socially beneficial in a more concentrated market. The Salant-Switzer-Reynolds model is a borderline case: $\lambda_i = 1$ for all i, so that the distribution of market shares among nonparticipants does not matter there.

Another interesting feature of condition (17) is that a merger is *more* likely to have a negative external effect, and hence to require careful antitrust scrutiny, when demand is *more* elastic. The reason is that with elastic demand outsiders' markups are small, so little welfare benefit is to be had from their increased output (while consumers still suffer from any price increase caused by the merger).

In the linear-quadrant model, unlike the constant marginal cost model, mergers may increase welfare even absent fixed-cost savings. In a previous version of this paper, we showed, by viewing infinitesimal mergers as transfers of small amounts of capital, that welfare rises ($\Delta W > 0$) if the merging firms are small and the rest of the industry is highly concentrated.[24] This is so even though there are no synergies, output falls, and concentration increases! Small merging firms have small price-cost margins, so a reduction in their output has little social cost, whereas large nonparticipant firms increase their output and this is socially valuable. In terms of condition (1), the increase in concentration, dH/H, outweighs the reduction in output, dX/X, yielding an *increase* in welfare. Thus, as hinted in the Introduction, structural changes such as mergers can lead to changes in H and in W that have the same sign, even without synergies. That this can happen more generally, with or without synergies, is shown by Proposition 5.

E. *Constant-Elasticity Demand and Constant Costs*

For constant marginal costs and constant-elasticity demand, $\lambda_i = 1 - s_i(1 + 1/\varepsilon)$. The net externality from an infinitesimal merger that raises price is posititve if and only if

$$2(s_1 + s_2) < 1 - \left(1 + \frac{1}{\varepsilon}\right)\sum_{i\in 0} s_i^2.$$

In this example, although the $p''' \geq 0$ condition of Proposition 5 does not hold, we can nonetheless show directly that any merger for which the above inequality holds at pre-merger shares generates a positive net externality. To establish this result, it is sufficient to show that the above inequality continues to hold as x falls. Differentiating totally, this is equivalent to $\Sigma_{i\in 0}\lambda_i\, ds_i \geq 0$. Since all the λ_i are positive, it is enough to show that s_i increases as X falls. But firm i's first-order condition requires that $s_i = \varepsilon(1 - c_x^i/p)$; since ε and c_x^i are constants, s_i must rise with p.

We therefore have derived, in an important special case, a sufficient condition in terms of the pre-merger market shares and the elasticity of demand, for every profitable merger to improve welfare. This result illustrates that the sufficient conditions of Proposition 5 are not necessary.

F. *Extending Beyond Cournot*

We have used the Cournot assumption in this section in two ways: (a) to measure outsiders' markups, $p - c_x^i$, by $-x_i p'(X)$, and (b) to compute outsiders' responses as measured by the λ_i's.[25] But within the class of homogeneous-goods models, our qualitative results extend well beyond Cournot behavior. Whatever the behavior of the nonparticipating firms, there are two external effects of a change in X_I. First, the price changes, causing a loss of $X_I dp$ to consumers.[26] Second, outsider firms are induced to change their outputs by some amounts, say $\psi_i dX_I$, and the social value of this is $\Sigma_{i \in O}[p - c_x^i]\psi_i dX_I$. Consequently, the externality is positive if and only if

---

[23] For example, if firm i has constant unit cost c_i and if $p'' > 0$, then λ_i will be inversely related to x_i.

[24] McAfee and Williams (1988) also construct examples of welfare-improving mergers in this model, using computer simulation.

$$(18) \quad \sum_{i \in O} \left[p - c_x^i\right] \psi_i dX_I - X_I dp > 0.$$

Equation (11) and Proposition 4 are a special case of equation (18). More generally, the ψ_i must be recalculated for each oligopoly theory, but the basic equation (18) persists.

To illustrate, we sketch here how our results would differ with nonzero conjectural variations. Suppose that firm i believes that its rivals will respond to its output changes with $dy_i/dx_i = v_i$.[27] Then firm i's equilibrium markup is $p - c_x^i = -x_i p'(X)(1 + v_i)$, and its equilibrium responsiveness is

$$\hat{\lambda}_i = -\frac{p' + x_i p''(1 + v_i)}{c_{xx}^i - p'(1 + v_i)}.$$

The condition in Proposition 4 for a positive externality becomes

$$\sum_{i \in O} \hat{\lambda}_i s_i (1 + v_i) > s_I.$$

We now ask how $\hat{\lambda}_i(1 + v_i)$ varies with v_i. This will tell us how the external effect of a small reduction in output by (say) firms 1 and 2 in an industry with a given vector of market shares $(s_1, s_2, \ldots, s_n)$, varies with the industry's competitiveness. In other words, *given* the observed market shares, how will the external effect of the merger depend upon the behavior that led to the observed pre-merger equilibrium? This is the relevant question for policy: it determines whether high or low perceived v_i's would make a proposed merger more socially acceptable, given the observed market shares.

Straightforward calculations show that $\hat{\lambda}_i(1 + v_i)$ is increasing in v_i if and only if

$$(1 + v_i)^2 x p' p'' - \left(p' + 2(1 + v_i) x p''\right) c_{xx} > 0.$$

This condition certainly holds if demand is linear and $c_{xx} \geq 0$: in that case, more collusive behavior (an increase in v_i) raises $\hat{\lambda}_i(1 + v_i)$, so an output reduction by firms 1 and 2 becomes more attractive to the rest of society, given all the firms' market shares. Perhaps surprisingly, the more tacit collusion that exists in pre-merger equilibrium and would exist after the merger (we assume the two are the same), the more likely it is that a privately profitable merger is socially desirable.

IV. Other Applications of the Externality Technique

The externality formula of Proposition 4 has applications that extend well beyond horizontal merger policy. We sketch two such applications here: (1) investment or commitment by an oligopolist, and (2) an import quota in an industry with a domestic oligopoly.

A. *Investment or Commitment in Oligopoly*

What is the welfare effect of a small, observable, unilateral change by an oligopolist that shifts outwards its reaction function? In particular, what external effects on consumers and other firms result from an oligopolist's decision to invest in new capital, or from an increase in its ability to commit to a high output level, for example, through long term contracts? As above, we can analyze these external effects in terms of an infinitesimal (but in this case a positive) change in X_I, where the subset I of "insiders" consists of the investing firm alone.[28] Applying Proposition 4 gives

---

[25] We also have used price to measure marginal gross benefits. This procedure is valid provided only that the good is homogeneous and buyers have no market power.

[26] The price effect on outsider firms' output, $X - X_I$, is merely a transfer from consumers to outsiders.

[27] The parameter v_i is called firm i's "conjectural variation." Although the conjectural variation model is logically flawed, it is a useful way of parameterizing the degree of competition among oligopolists. Such behavior might, for example, arise as the equilibrium of an (unmodeled) dynamic oligopolistic game.

PROPOSITION 6: *The external effect of a small outward shift in firm 1's reaction function has the sign of*

$$s_1 - \sum_{i=2}^{n} \lambda_i s_i. \tag{19}$$

Expression (19) is most likely to be negative if firm 1's rivals have large market shares and therefore high price-cost margins, and if their equilibrium outputs are sensitive to expansion by firm 1. In such a market, firm 1 "steals" a great deal of valuable business from its rivals when it invests. This negative externality imposed on rivals must be compared against the transfer to consumers of the price change on firm 1's output.[29] Clearly, if firm 1 is sufficiently small, (19) is negative. Thus, although investment by a small firm lowers price and benefits consumers, it harms rival firms by more.

The impact on rivals and on consumers of allowing one firm in a Cournot oligopoly some power of pre-commitment is given also by Proposition 6. If we move from a régime of less commitment power to one of more, we have $dX_I > 0$. Consequently, the external welfare effects of allowing more output commitment by one firm are beneficial if and only if that firm has a sufficiently large market share. And it is certainly profitable for the firm with the commitment power. Hence, if one firm has a market share in Cournot oligopoly that exceeds the λ-weighted sum of all others', then it is socially desirable to give that firm some Stackelberg power. And if the conditions of Proposition 5 hold, it follows that it is desirable to go all the way to give the firm complete Stackelberg power.

As an illustration, take the simplest possible model. There are two firms, each with zero marginal costs, and $p(X) = 1 - X$. Recall that $\lambda_i = 1$ with this specification, so that, in the (symmetric) Cournot equilibrium, $s_1 = \frac{1}{2} = \lambda_2 s_2$. In Cournot equilibrium, each firm produces output 1/3, and makes profits of 1/9, while consumer surplus is 2/9. But if firm 1 is Stackelberg leader, then $x_1 = 1/2$ and $x_2 = 1/4$, giving firm 2 profits of 1/16 and consumers surplus of 9/32. Since $(9/32 + 1/16) > (2/9 + 1/9)$, firm 2 and consumers jointly benefit from the shift to Stackelberg behavior, as Proposition 5 predicted. Of course, firm 1 also gained.

B. *Import Quotas in Oligopolistic Industries*

We now show how Propositions 4 and 5 illuminate the domestic welfare consequences of quotas when the domestic industry consists of a Cournot oligopoly (plus perhaps a competitive fringe). Thinking of foreign suppliers as the "insiders," we can view a tightening of a quota as a reduction in X_I. Moreover, the welfare of outsiders (domestic firms) and consumers is simply domestic welfare. Proposition 4 therefore gives us:

PROPOSITION 7: *Consider the effect of an import quota in an industry where domestic firms are Cournot oligopolists, perhaps including a competitive frings. Slightly tightening the quota raises domestic welfare if and only if the share of imports is less than the λ-weighted sum of domestic producers' shares.*

Proposition 5—when its assumptions hold—implies that a small enough import sector should be excluded altogether![30] Although this would hurt consumers, it would help domestic producers by more. For instance, in the constant-cost, linear-demand case, Proposition 5 states that if imports have no more than half the market in Cournot equilibrium, domestic welfare would be greater if imports were completely excluded! To illustrate, take the same model as in the previous subsection: demand is $p=1-X$ and there are two firms (one now foreign), each with zero marginal costs. Domestic welfare in Cournot equilibrium is $2/9+1/9=1/3$. Excluding the foreign firm creates a domestic monopoly, which will produce output of $1/2$, yielding (domestic) welfare of $(1/8+1/4)=3/8>1/3$.

---

[28] Additional applications of Proposition 4 to investment are presented in our (1990) paper.

[29] The price effect on other firms' outputs is merely a transfer from rivals to consumers. Marius Schwartz (1988) discusses some similar examples in a special model.

[30] Janusz Ordover and Robert Willig (1986) obtain a similar result in a model with linear demand and constant unit costs.

V. Conclusions

We have studied the output and welfare effects of mergers in Cournot oligopoly, using cost and demand functions restricted only by standard stability assumptions.

We found that mergers do indeed, typically, raise price. In particular, any merger that generates no synergies, in the sense of equation (8), raises price. And for a merger to lower price requires considerable economies of scale or learning. This result is only strengthened if the merger causes behavior in the industry to shift from Cournot to something less competitive. Furthermore, and consistent with the approach taken by the merger guidelines, we found that the economies of scale or learning effects necessary for a merger to lower price are greater, the larger are the merging firms' market shares and the less elastic is industry demand.

In our policy analysis, we focused on the external effects of a merger rather than trying directly to sign its overall welfare effect. This approach, as well as being consistent with market-oriented policy analysis in general, has a great practical advantage: the information required is much more readily available. Looking at the external effect would also allow antitrust authorities to make use of the fact that only privately profitable mergers are proposed, and it would permit them to give cost savings a larger role in merger policy (as Williamson's (1968) calculations suggest they deserve) without demanding either information or credulity about alleged synergies.

The information that *is* needed in our analytical scheme is in two separable parts. First, to test whether a merger will indeed reduce output requires information only on participants' (pre-merger) marginal-cost functions and on that of the merged entity. Although this information may be hard to obtain, our observation does indicate exactly what the relevant information is. Information about market demand and other firms' costs is not relevant for determining whether price will rise or fall, so long as Cournot behavior applies both before and after the merger. Second, to sign the external effect of an output-reducing merger—whether it benefits or harms rival firms and consumers jointly—requires information only on market shares and the output responsiveness parameters (the λ_i) of nonparticipants.

Inevitably, for general results one requires information not only at the pre-merger equilibrium, but along a path from pre-merger to post-merger equilibrium. But Proposition 5 gives conditions—general enough to cover the special cases previously studied and many more—under which no such global inquiry is required: one need merely compare the participants' pre-merger joint market share with the λ-weighted sum of their rivals'. We can thus give surprisingly general (sufficient) conditions under which all privately profitable mergers raise welfare, and presumably all proposed mergers should be approved.

Our techniques have a wide range of application in oligopoly theory, as Propositions 6 and 7 show. Quite generally, our results imply that many output-reducing changes should not be grudgingly tolerated, but should be positively welcomed by the rest of society. The counterintuitive, almost paradoxical, fact is that, in Cournot and similar theories of oligopoly with homogeneous goods, the presence of small firms with little market power is not desirable. Their output, produced at a marginal cost that almost consumes its gross social benefit, also displaces or discourages output at (larger) firms with lower marginal costs (this displacement effect is, of course, absent in perfect competition). Consequently, it often enhances economic welfare—defined in the usual way—to close down small or inefficient firms, or, failing that, to encourage them to merge so that they produce less output. This observation may call for some rethinking of our views on policy toward competition, including horizontal merger policy.

We believe that our paper provides useful tools and insights for horizontal merger policy. But our analysis is limited in two major ways, both of which are inherent in the standard Cournot model (and its conjectural-variations extensions). First, we analyze a homogeneous-goods industry; our results may not apply well to markets in which product differentiation is substantial. Applying techniques such as those presented here to differentiated-product industries is an important topic for future research. Second, by assuming that the firms behave as Cournot competitors, both before and after the merger, we ignore any effect of a merger on the probability or nature of explicit collusion. We believe that the Cournot model captures the notion of *tacit* collusion fairly well, but we are well aware that many antitrust practitioners are more concerned with the possibility of *explicit* collusion than with the nature of tacit collusion. Unfortunately, there is no fully satisfactory model of the probability of successful explicit collusion in oligopoly. Such a model could be used to estimate a merger's effect on the probability of collusion.

APPENDIX

The Herfindahl Index in Cournot Oligopoly

Consider a Cournot oligopoly in which firm i ($i = 1,\dots,n$) produces output x_i at cost $c^i(x_i)$. Total output is $X \equiv \Sigma_{i=1}^n x_i$, and the market price, $p \equiv p(X)$, measures the marginal gross benefit of output. For an arbitrary change $\{dx_i\}$ in firms' outputs,

$$dW = \sum_{i=1}^{n} \left(p - c_x^i\right) dx_i .$$

In Cournot equilibrium, firm i's first-order condition gives $p - c_x^i = -p'(X)x_i$. Hence,

$$dW = -p'(X) \sum_{i=1}^{n} x_i \, dx_i .$$

Now this change in welfare is related to the Herfindahl index, H, since

$$\sum_{i=1}^{n} x_i \, dx_i = \frac{1}{2} d\left[\sum_{i=1}^{n} x_i^2\right] = \frac{1}{2} d\left[X^2 H\right] = XH\,dX + \frac{1}{2} X^2 \, dH .$$

Thus we have

$$dW = -X^2 H p'(X) \left\{ \frac{dX}{X} + \frac{1}{2}\frac{dH}{H} \right\}.$$

With downward-sloping demand, $p'(X) < 0$, so dW has the sign of

$$\frac{dX}{X} + \frac{1}{2}\frac{dH}{H} .$$

See Keith Cowling and Michael Waterson (1976) or Robert Dansby and Robert Willig (1979) for related calculations showing the welfare significance of the Herfindahl Index in Cournot oligopoly.

Proof of Lemma

Write Δx_1 and ΔX for the changes in firm 1's output and in aggregate output. It is enough to prove the Lemma for an infinitesimal change dx_1, since $0 < dX/dx_1 < 1$ implies $0 < \Delta X/\Delta x_1 < 1$. For any firm $i \neq 1$, $dx_i = -\lambda_i dX$. Adding up for $i \neq 1$ we have $dy_1 = -\Sigma_{i \neq 1} \lambda_i \, dX$. Adding dx_1 to this equation gives $dX = -\Sigma_{i \neq 1} \lambda_i \, dX + dx_1$, or $dX(1 + \Sigma_{i \neq 1} \lambda_i) = dx_1$. With conditions (3) and (4), each λ_i is positive, so dX has the same sign as dx_1 but is smaller in magnitude. □

Proof of Proposition 1

By our Lemma, we need only sign the effect on the insiders' total output. And in order to do that, we need only find whether the new firm M would increase or decrease its output if the nonparticipant firms held their outputs constant at pre-merger levels.

Denote pre-merger outputs by $\bar{x}_i$ and $\bar{X}$, and call the insiders' aggregate pre-merger output $\bar{X}_M$. At pre-merger output levels, M's marginal revenue is $p(\bar{X}) + \bar{X}_M p'(\bar{X})$. So M will reduce its output if and only if $c_x^M(\bar{X}_M) \geq p(\bar{X}) + \bar{X}_M p'(\bar{X})$, or $p(\bar{X}) - c_x^M(\bar{X}_M) \leq -\bar{X}_M p'(\bar{X})$. Thus it is enough to show that $-\bar{X}_M p'(\bar{X})$ is equal to the sum of the merging firms' pre-merger markups. But, for each firm $i \in I$, $c_x^i(\bar{x}_i) = p(\bar{X}) + \bar{x}_i p'(\bar{X})$, or $p(\bar{X}) - c_x^i(\bar{x}_i) = -\bar{x}_i p'(\bar{X})$. Adding this up over $i \in I$, we find that

$$\sum_{i \in I} \left[p(\bar{X}) - c_x^i(\bar{x}_i) \right] = -\bar{X}_M p'(\bar{X}),$$

as required. □

Proof of Proposition 2

We give the proof for a two-firm merger, but it extends to multi-firm mergers. The proof proceeds in four steps.

Step 1. First, by the Lemma, the change in total output has the same sign as the change in the insiders' output. Therefore, it is enough for us to show that the insiders will want to reduce their aggregate output, if the outsiders' output is held constant at $\bar{X}_O$. For notational simplicity, since the outsiders' output is fixed, we shall write $p(x_1+x_2)$ instead of $p(x_1+x_2+\bar{X}_O)$.

Step 2. Next, we show by a revealed-preference argument that both firms' outputs cannot rise when the two firms maximize joint profits rather than their own profits (as in Cournot equilibrium).

Denote by $\bar{x}_i$ firm i's output in pre-merger Cournot equilibrium, and by x_i its output when the two firms maximize joint profits. We will show that it is impossible to have $x_i > \bar{x}_i$ for $i=1,2$. By revealed preference,

$$(20)\quad (x_1+x_2)p(x_1+x_2)-c^1(x_1)-c^2(x_2) \geq (\bar{x}_1+\bar{x}_2)p(\bar{x}_1+\bar{x}_2)-c^1(\bar{x}_1)-c^2(\bar{x}_2),$$

and

$$(21)\quad \bar{x}_1 p(\bar{x}_1+\bar{x}_2)-c^1(\bar{x}_1) \geq x_1 p(x_1+\bar{x}_2)-c^1(x_1),$$

$$(22)\quad \bar{x}_2 p(\bar{x}_1+\bar{x}_2)-c^2(\bar{x}_2) \geq x_2 p(\bar{x}_1+x_2)-c^2(x_2).$$

Adding equations (21) and (22), and comparing to (20), yields

$$(23)\quad (x_1+x_2)p(x_1+x_2) \geq x_1 p(x_1+\bar{x}_2)+x_2 p(\bar{x}_1+x_2),$$

which is inconsistent with both $x_1 \geq \bar{x}_1$ and $x_2 \geq \bar{x}_2$ unless both hold with equality. Without loss of generality, suppose that the output at facility 2 does not rise after the merger: $x_2 \leq \bar{x}_2$.

Step 3. Let x_1' be the output that maximizes π_1 given outputs x_2 and $\bar{X}_O$. Since $x_2 \leq \bar{x}_2$, the Lemma tells us that $x_1' \geq \bar{x}_1$ and $x_1' + x_2 \leq \bar{x}_1 + \bar{x}_2$.

Step 4. Finally, if $x_2 > 0$, when firm 1 is maximizing joint profits, it will produce strictly less (given firm 2's output) than if it were maximizing only its own, since increases in x_1 reduce the profits earned at facility 2. That is, $x_1 < x_1'$. Using Step 3, we thus have $x_1 + x_2 < \bar{x}_1 + \bar{x}_2$ as was required. If $x_2 = 0$ then $x_1 = x_1'$ and the result follows directly from the Lemma. □

Economies of Scale

For equilibrium output to increase, Proposition 1 requires that marginal cost at twice the pre-merger capital and output levels must be less than pre-merger marginal cost, by an amount equal to the pre-merger markup. That is,

$$c_x(2x,2k) \leq c_x(x,k) - [p - c_x(x,k)],$$

where p is the pre-merger price and x is each firm's pre-merger output. Now the pre-merger markup $p - c_x(x,k)$ is equal, by the pre-merger first-order condition, (2), to $s/(\varepsilon - s)$ times pre-merger marginal cost $c_x(x,k)$. So output falls unless

$$c_x(2x,2k) \leq \left[1 - \frac{s}{\varepsilon - s}\right] c_x(x,k).$$

If $c(x,k)$ is dual to the generalized Cobb-Douglas production function $f(L,k) = k^a L^b$, then $c_x(x,k) \equiv w/b x^{1/b-1} k^{-a/b}$, where w is the price of the variable input L, so that

$$\frac{c_x(2x,2k)}{c_x(x,k)} = 2^{1/b-1-a/b}.$$

Hence, (9) states that

$$a + b \geq 1 + b\log_2(\varepsilon - s) - b\log_2(\varepsilon - 2s).$$

If $s = 0.2$ and $\varepsilon = 1$, and if $a = b$, this requires $a \geq 1/\log_2 3$.

Learning. Suppose that firms 1 and 2 merge, capital is immobile and marginal costs are nondecreasing. We prove here that for price to fall, the merger must either reduce θ_1 by at least a factor $s_2/(\varepsilon - s_1)$ or reduce θ_2 by at least a factor $s_1/(\varepsilon - s_2)$.

At least one of the insiders, call it firm 1, must at least maintain its pre-merger output level if price is not to rise. And, by the envelope theorem, M's marginal cost will equal firm 1's marginal cost at its post-merger level of output. But, by equation (7), for output to rise we must have $c_x^1 - c_x^M > p - c_x^2$, so firm 1's marginal costs must fall by at least firm 2's markup. However, with $c_{xx}^1 \geq 0$ and firm 1 expanding, firm 1's marginal costs could fall only due to synergies. If price is to fall, therefore, synergies must reduce firm 1's marginal costs by at least the extent of firm 2's markup.

Just how strong a synergistic cost reduction is needed for this condition to be met and thus reverse the presumption that aggregate output will fall with the merger? In the pre-merger equilibrium, we have $(p - c_x^i(x_i))/p = -x_i p'/p = s_i/\varepsilon$. Hence, $(p - c_x^2)/p = s_2/\varepsilon$ and $p/c_x^1 = \varepsilon/(\varepsilon - s_1)$.

The reduction in firm 1's marginal costs required for price to fall is $c_x^1 - c_x^M > p - c_x^2$. In percentage terms, this requires

$$\frac{c_x^1 - c_x^M}{c_x^1} > \frac{p - c_x^2}{p}\,\frac{p}{c_x^1}.$$

Substituting for the two factors on the right-hand side, we must have

$$\frac{c_x^1 - c_x^M}{c_x^1} > \frac{s_2}{\varepsilon}\,\frac{\varepsilon}{\varepsilon - s_1} = \frac{s_2}{\varepsilon - s_1} = \frac{p}{c_x^1}.$$

Learning must reduce firm 1's marginal costs by at least $s_2/(\varepsilon - s_1)$ for price to fall. □

Proof of Proposition 5

We will show that the conditions given imply that $d[\lambda_i x_i]/dX \leq 0$ for $i \in O$. Since an output-reducing merger involves a reduction in X_I, an infinitesimal merger's effect on $\eta \equiv \Sigma_{i \in O} \lambda_i x_i - X_I$ is therefore unambiguously positive. So, after doing an infinitesimal merger that benefits nonparticipants, we will find that a further infinitesimal merger benefits them by even more, and so on until the merger is complete. Evidently, this implies (it is much stronger than) Proposition 5.

Recall that we are considering an exogenous change dX_I, which induces changes dx_i by outside firms i, and hence induces a change dX in aggregate output and changes $d\lambda_i$ in the λ's. Since $d[\lambda_i x_i] = \lambda_i dx_i + x_i d\lambda_i$, and since $dx_i = -\lambda_i dX$, we have

$$d[\lambda_i x_i] = -\lambda_i^2 dX + x_i d\lambda_i.$$

Now think of $\lambda_i \equiv -p'(X) + x_i p''(X)/c_x^i(x_i) - p'(X)$ as a function of two variables, X and x_i. Thus we have

$$\frac{d[\lambda_i x_i]}{dX} = -\lambda_i^2 + x_i \left(\frac{\partial \lambda_i}{\partial X} - \lambda_i \frac{\partial \lambda_i}{\partial x_i} \right).$$

Substituting in for λ_i and its partial derivatives, we get

$$\begin{aligned} (24) \quad (c_{xx} - p')^2 \frac{d[\lambda_i x_i]}{dX} &= -(p' + x_i p'')^2 - x_i^2 p'''(c_{xx} - p') \\ &\quad + x_i c_{xxx} \frac{(p' + x_i p'')^2}{c_{xx} - p'} \\ &\quad - x_i p''[c_{xx} + p' + 2x_i p'']. \end{aligned}$$

For quadratic $p(\cdot)$ and $c(\cdot)$ functions, we can ignore the terms in c_{xxx} and in p''', and then $d[\lambda_i x_i]/dX$ has the sign of

$$\begin{aligned} &-x_i p''[c_{xx} + p' + 2x_i p''] - (p' + x_i p'')^2 \\ &\equiv -x_i p'' c_{xx} - \left(p' + \frac{3}{2} x_i p'' \right)^2 - \frac{3}{4} x_i^2 (p'')^2, \end{aligned}$$

which is negative provided that p'' and c_{xx} are nonnegative (actually, provided that their product is nonnegative). The proposition follows by inspection of equation (24), since p''' and c_{xxx} enter with unambiguous signs. □

REFERENCES

Cowling, Keith and Waterson, Michael, "Price-Cost Margins and Market Structure," *Economica*, August 1976, *43*, 267–74.

Dansby, Robert E. and Willig, Robert D., "Industry Performance Gradient Indexes," *American Economic Review*, March 1979, *69*, 249–60.

Demsetz, Harold, "Industry Structure, Market Rivalry, and Public Policy," *Journal of Law and Economics*, April 1973, *16*, 1–10.

______, "Two Systems of Belief About Monopoly," in Harvey Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston, MA: Little Brown, 1974, 164–84.

Deneckere, Raymond and Davidson, Carl, "Incentives to Form Coalitions with Bertrand Competition," *Rand Journal of Economics*, Winter 1985, *16*, 473–86.

Dixit, Avinash, "Comparative Statics for Oligopoly," *International Economic Review*, February 1986, *27*, 107–22.

Farrell, Joseph and Shapiro, Carl, "Horizontal Mergers: An Equilibrium Analysis," John M. Olin Program for the Study of Economic Organization and Public Policy, Discussion Paper No. 17, Princeton University, June 1988.

______ **and** ______, "Asset Ownership and Market Structure in Oligopoly," *Rand Journal of Economics*, March 1990, *21*, forthcoming.

Fisher, Franklin M., "Horizontal Mergers: Triage and Treatment," *Journal of Economic Perspectives*, Fall 1987, *1*, 23–40.

Gaudet, Gérard and Salant, Stephen, "The Profitability of Exogenous Output Contractions in Cournot Equilibrium: Implications for Strikes, Mergers and Export Subsidies," University of Michigan, May 1988.

McAfee, R. Preston and Williams, Michael A., "Horizontal Mergers and Antitrust Policy," unpublished manuscript, U.S. Department of Justice, February 1988.

McElroy, F. William, "Price and Welfare Effects of Oligopolistic Mergers and Partial Acquisitions," unpublished manuscript, Georgetown University, November 1988.

Ordover, Janusz A. and Willig, Robert D., "Perspectives on Mergers and World Competition," in R. Grieson, ed., *Antitrust and Regulation*, Lexington, MA: Lexington Books, DC Heath, 1986, 201–18.

Perry, Martin K. and Porter, Robert H., "Oligopoly and the Incentive for Horizontal

Merger," *American Economic Review*, March 1985, *75*, 219–27.

Salant, Stephen, Switzer, Sheldon and Reynolds, Robert, "Losses Due to Merger: The Effects of an Exogenous Change in Industry Structure on Cournot-Nash Equilibrium," *Quarterly Journal of Economics*, May 1983, *98*, 185–199.

Schwartz, Marius, "Investments in Oligopoly: Welfare Effects and Tests for Predation," *Oxford Economic Papers*, 1989, *41*, forthcoming.

Shapiro, Carl, "Theories of Oligopoly Behavior," in R. Schmalensee and R. Willig, eds., New York: *The Handbook of Industrial Organization*, Elsevier North-Holland, 1989.

U.S. Department of Justice, "Merger Guidelines," 2 Trade Reg. Rep. (CCH), 4491–95, 1984.

White, Lawrence J., "Antitrust and Merger Policy: A Review and Critique," *Journal of Economic Perspectives*, Fall, 1987, *1*, 13–22.

Williamson, Oliver E., "Economies as an Antitrust Defense: The Welfare Tradeoffs," *American Economic Review*, March 1968, *58*, 18–36.

Equilibrium Vertical Foreclosure

By JANUSZ A. ORDOVER, GARTH SALONER, AND STEVEN C. SALOP*

We formulate a complete, but analytically simple, equilibrium model of vertical mergers to evaluate the logic of standard vertical foreclosure claims and the criticisms made of those claims. The model includes incentives of the integrated firm and unintegrated input suppliers to exclude rivals, the potential counter-strategies of competitors to these foreclosure threats, and the potential hold-out problem. In this fully specified model, vertical foreclosure can emerge in equilibrium. (JEL 610)

The competitive effects of vertical mergers have long been a source of controversy in economics and antitrust. One of the central issues in that debate involves vertical foreclosure; the exclusion that results when unintegrated downstream rivals are foreclosed from the input supplies controlled by the firm that integrates. Analogous effects occur when unintegrated upstream competitors are foreclosed from selling to the downstream division of the integrated firm.[1]

While foreclosure has been accepted in leading court decisions[2] and policy guidelines,[3] critics maintain that the theory itself is logically flawed. They claim that a vertically integrated firm will have no incentive to exclude its rivals, and if it did try to exclude them, rivals could protect themselves by contracting with other unintegrated firms.[4]

This controversy can be seen more clearly by making the vertical foreclosure theory more specific.[5] According to the theory, a single vertical merger can disadvantage downstream rivals as follows.[6] Consider a market in which the supply of inputs is competitive before the merger and there are no production efficiency benefits gained from vertical integration. After the merger, suppose the upstream division of the now-integrated firm refuses to supply inputs to the rivals of its downstream division.

This foreclosure of rivals from these supplies means that remaining suppliers will face less competition. As a result, they may be able to increase their profits by raising their input prices to the unintegrated down-

---

*Department of Economics, NYU; Department of Economics and Sloan School of Management, MIT; and Georgetown University Law Center, respectively. We have benefited from several useful discussions with Mike Whinston and comments from William Comanor, Louis Kaplow, Julio Rotemberg, Jean Tirole, an anonymous referee, and participants at a conference on Vertical Restraints at Harvard University, September 1987. Financial support from the National Science Foundation (grant IST-8510162), the C.V. Starr Center for Applied Economics at NYU, the Hoover Institution at Stanford University, and the Sloan Foundation is gratefully acknowledged.

[1] The other main strand of the debate concerns how vertical mergers can eliminate inefficiencies due to a double markup of input costs and substitution to other less efficient inputs. Similar models show how vertical mergers can be used to facilitate price discrimination against purchasers with less elastic demands. This literature is reviewed by Martin Perry (1989) and Jean Tirole (1988).

[2] See, for example, *Brown Shoe Co. v. United States*, 370 US 294 (1962) and *United States v. Aluminum Co. of America*, 44 F. Supp. 97, 121-44 (SDNY, 1941), 148 F. 2d 416, 436-38 (2d Cir., 1945).

[3] For example, the Department of Justice's 1984 Vertical Merger Guidelines.

[4] See Robert Bork (1978), pp. 222–45 and pp. 299–309, Richard Posner (1976), pp. 171–84, 196–207, Frank Easterbrook (1984), and Posner and Easterbrook (1981), pp. 869–76.

[5] In what follows, we focus on foreclosure of downstream rivals. The analysis of foreclosure of upstream rivals is analogous.

[6] For a general discussion of potential anticompetitive consequences of vertical mergers see Thomas Krattenmaker and Steven Salop (1986).

stream firms.[7] These higher prices benefit the vertically integrated firm. If rivals' costs of inputs are increased, they will be forced to reduce their production and raise the prices they charge in the downstream market. This reduction in competition allows the downstream division of the integrated firm to increase its market share and its price. Thus the profits of the vertically integrated firm can rise, even if there are no production efficiency benefits flowing from the vertical integration.

The critics of vertical foreclosure raise a number of objections to these assertions. Each of these objections implies that the strategy of vertical integration and foreclosure of supply will not increase the overall profitability of the integrated firm. The objections deny the ability to use power in one market to "leverage" into a position of power in a second market. For example, Robert Bork (1978, p. 231) concludes:

> Vertical merger does not create or increase the firm's power to restrict output. The ability to restrict output depends on the share of the market occupied by the firm. Horizontal mergers increase market share, but vertical mergers do not.

There are six main objections to the foreclosure theory. First, the supply of inputs available to rivals is not necessarily reduced as a result of a vertical merger. This is because the integrated firm also reduces its demand for inputs produced by unintegrated suppliers. For example, if a firm with 10 percent of the output market purchases a supplier that accounts for 10 percent of the input market, it will not reduce the net supply available to rivals. Instead, it merely will necessitate a rearrangement in supply relationships.

Second, critics assert that it may not be profitable for the integrated firm to foreclose its downstream rivals. Although the profits of the downstream division rise, the upstream division of the integrated firm loses input sales from its refusal to deal with rival downstream firms. These lost upstream profits may exceed the increased downstream profits. As a result, it does not follow necessarily that it is in the interest of the vertically integrated firm to foreclose downstream rivals.

Third, remaining suppliers may not have the incentive to raise their input prices. The foreclosure theory relies on the incentive of remaining suppliers to increase their input prices after they no longer face competition in the input market from the integrated supplier. But this reduction in competition is more apparent than real. If the remaining suppliers raise their input prices, downstream firms often can begin to produce the input themselves.

Even if such entry is assumed to be impossible, raising prices may still be unprofitable. If the remaining suppliers raise prices, their downstream customers will be placed at a cost disadvantage vis-à-vis the integrated firm and will be forced to reduce their input purchases. Stated more directly, unintegrated firms' input demands are highly elastic. Consequently, suppliers' incentives to increase prices even after the integrated firm forecloses those rivals from its input production are low.

Fourth, the foreclosed rivals have available a number of alternative counterstrategies to battle their exclusion. In particular, they can respond by vertically integrating with the remaining unintegrated input producers. This will enable them to obtain the inputs at the competitive price, thereby eliminating their cost disadvantage. Thus, the first firm to integrate will gain no market power.

Fifth, it is not obvious that the first downstream firm's bid to integrate with an upstream firm will succeed to begin with. The vertical foreclosure theory suffers from a potential holdout problem. Assuming that remaining input producers gain the power to increase their input prices to unintegrated downstream buyers, a holdout supplier would gain an advantage by remaining unintegrated. Unless the downstream firm bids enough to compensate its potential merger partner for this lost opportunity, its bid will fail. Each firm will hold out, in the anticipation that another supplier will merge and give it the ability to profitably raise its input price. If the downstream firm does try to compensate its potential merger partner for this lost opportunity, the profitability of the merger is decreased, possibly to the point that no merger will occur.

---

[7] Krattenmaker and Salop term this version of the vertical foreclosure theory the "Frankenstein Monster" because the integrated downstream firm creates market power for the remaining upstream firms, giving them the power to raise input prices to its rival.

Finally, the vertical foreclosure theory suffers from one last potential flaw. Since the firm that is foreclosed is placed at a disadvantage, it ought itself to participate in the bidding for the scarce upstream resource. While this flaw is somewhat different from the earlier criticisms, in that it does not amount to a claim that vertical foreclosure will fail, it does make the theory incomplete.

In this paper, we set out a complete, but analytically tractable equilibrium model to evaluate the logic of the standard vertical foreclosure argument as well as the arguments advanced by its critics. Even though our model is simple, it includes the incentives of the integrated firm and remaining unintegrated input suppliers to exclude rivals, the possible counterstrategies of competitors to these threats of foreclosure, and the potential holdout problem. It also deals explicitly with the bidding for the scarce upstream resource.

We analyze the simplest vertical structure capable of capturing the essential elements of interest. In particular, we consider successive duopoly: two upstream firms initially provide an equal share of a homogeneous input to two downstream producers. These two downstream firms each produce a differentiated product in competition with the other, and initially have equal market shares. In order to focus exclusively on foreclosure rather than efficiency, we have structured the model so that double marginalization does not arise when the firms initially are unintegrated. This is accomplished by assuming Bertrand competition upstream.

We then allow for the possibility that a downstream firm acquires one of the upstream suppliers, leaving the rival with a single independent upstream supplier. However, the net supply available to the rival does not necessarily fall. Of course, if the integrated firm refuses to supply the unintegrated downstream firm, the remaining upstream supplier is endowed with a good deal of market power. We do not assume, however, that the integrated firm will refuse to deal with the remaining downstream rival. Rather, we examine several different scenarios for the way in which the input prices are determined. We also examine the possibility that the remaining downstream rival is able to counter the effect of the first vertical merger by itself merging with the remaining supplier. Finally, we require that the integrating downstream firm outbid its rival for the acquisition of the upstream firm, that is, we require bidding for the right to foreclose.

In this fully specified model, vertical foreclosure does emerge in equilibrium, provided the downstream firms' revenues are increasing in the price of the input. Then, when one of the firms integrates backward, thereby causing an increase in the input prices of the rival, downstream prices and revenues increase. Of course the profits of the unintegrated downstream firm decrease because it must pay the higher input costs. However, the sum of its profits and those of its supplier (which sum equals downstream revenues) increase. Therefore its loss in profits is less than the increase in the profits of its supplier. Accordingly, the unintegrated downstream firm cannot induce the upstream firm to merge and return to the *ex ante* position.

This argument holds for small increases in the upstream price. If the unintegrated upstream firm has an incentive to raise its price excessively, however, the unintegrated downstream firm will lose so much market share to its rival that its revenue will fall. Then the unintegrated firms will have an incentive to merge.

Thus, whereas the conventional critique of foreclosure theory has questioned the ability of the remaining upstream firm(s) to raise price, ironically, in this model a necessary condition is that it not be able to raise its price too much. Its ability to so raise the price will be reduced if it faces competition from other substitutes, or, as seems plausible, if the integrated firm is able to "keep a lid" on that price by announcing a price of its own at which it is prepared to supply the rival downstream firm.

The condition that the downstream firms' revenues be increasing in the price of the input tends to hold in the case of differentiated product price competition. There if a firm increases its price in response to an increase in the price of an input (an action the rival likes), the rival tends to respond in like manner, raising its own price: prices are strategic complements. In the case of Cournot competition with downward sloping best-response functions, in contrast, the firms' actions are strategic substitutes: if a firm reduces its output in response to an increase in the input price (an action the rival likes), the rival responds by increasing its own output, leading to a decrease in the former's revenues. Thus, while the analysis rests on certain features of the market structure, they are not the ones suggested by the conventional critique.

There is an earlier literature on vertical foreclosure which focuses on the case in which either the upstream or downstream tier of the industry is monopolized (or, at least, one firm has a dominant market position). That literature is concerned with the question of whether the dominant firm can increase its profitability by foreclosing some of the firms at the other level.[8] The fundamental difference here is our focus on vertical foreclosure by a firm to increase its market power in the level of the market which it currently shares with another rival.

Our work is related to that of Michael Salinger (1988). He examines whether vertical integration between successive oligopoly leads to higher or lower final goods prices. His model is more general than ours in that he allows for successive oligopoly rather than duopoly, and he allows for differing numbers of upstream and downstream firms. A similar body of work (focusing mainly on the duopoly case) also has been developed in the marketing literature.[9] However, none of this work examines the issues of counterstrategies and bidding which are the focus of our paper.

The paper is organized as follows: We begin by laying out the basic model in Section I. That section also establishes the unintegrated and fully integrated outcomes as benchmarks for our further analysis. In Section II, we analyze the model of vertical integration with differentiated products and establish the rationality of foreclosure in that setting.

In Section III, we study the robustness of our conclusions to changes in model specification. We show that the holdout problem may arise when upstream firms bid for downstream firm rather than vice versa because downstream firms earn positive profits before the merger. We also show that merger incentives depend on the nature of competition among the downstream firms. In particular, when downstream firms compete in quantities there are no foreclosure-creating vertical mergers in equilibrium because it is always profit-increasing for the unintegrated firms to merge in response to the initial vertical merger. Finally, we consider the equilibria that can be implemented with two-part tariffs for inputs in lieu of a vertical merger. We argue that a two-part tariff contract between an input supplier and a buyer that does not specify prices to other buyers will not implement the anticompetitive equilibrium with foreclosure identified in Section II. Section IV concludes.

---

[8] See William Comanor (1967) for an early example of a paper in this line, and Jean Tirole (1988, pp. 193–4) for a survey of the issues. Tirole (1988, ch. 4) also provides an excellent survey of vertical restraints such as "exclusive dealing" (as discussed, for example, in Comanor and H. E. Frech, 1985, Frank Mathewson and Ralph Winter, 1987, and Marius Schwartz, 1987), and "exclusive territories" (as discussed, for example, in Patrick Rey and Joseph Stiglitz, 1986). The paper by Schwartz points to the importance of considering a retailer's rationality in accepting an exclusive dealing arrangement involving a higher wholesale price, while Mathewson and Winter emphasize the importance of considering competition among wholesalers for the right to be selected by the retailer. Those papers underscore the importance of considering a rich set of strategies and counterstrategies of the kind considered in the current paper. In a somewhat different vein, Philippe Aghion and Patrick Bolton (1987) show how a firm with a first-mover advantage can deter entry by entering into an exclusive contract with customers.

[9] Two good examples are Sridhar Moorthy (forthcoming) and Ann Coughlin and Birger Wernerfelt (1988).

I. The Industry Structure and Benchmark Outcomes

In this section we lay out the structure of the market setting we examine and illustrate how the standard foreclosure argument can be applied to it. In the following section, we enrich the setting to model merger strategies, price-setting of the upstream good, and merger counterstrategies.

A. *Basic Model*

Two upstream firms, $U1$ and $U2$, provide a homogeneous input to two downstream firms, $D1$ and $D2$. For reasons that will become obvious soon, we assume that the input is produced with a constant returns to scale technology. At this point, we do not make any assumptions regarding the production technology for the outputs of $D1$ and $D2$.

Throughout the paper we assume static competition between the downstream firms and concentrate on the case where the firms produce differentiated products and use prices as their strategic variables. We denote by $Q^i(p_i, p_j)$, $i, j = 1,2$, the demand function for the product of firm i.

B. *Vertical Relationships*

There are numerous ways in which vertical relationships can be structured. In order to establish benchmark outcomes, it is useful to analyze two polar cases, (i) where neither downstream firm is vertically integrated (the "stand-alone" case); and (ii) where both are integrated and the internal transfer price is efficiently set at marginal cost. Where neither firm is vertically integrated, the input prices are determined by the nature of upstream competition. In the analysis of the benchmark case, we assume that when the stand-alone case obtains, the input suppliers compete in prices. Thus, Bertrand competition between $U1$ and $U2$ drives the input price to marginal cost.

Between these two extremes lie a variety of alternative pricing mechanisms. Of special interest to us will be those that can obtain when only one of the firms is vertically integrated. Then, the unintegrated firms may use long-term contracts to set nonlinear prices (such as two-part tariffs), for example, while an integrated firm may be able credibly to commit to internal transfer prices different from marginal cost.

C. *Equilibrium in the Downstream Market*

$D1$ and $D2$ engage in Bertrand competition which yields equilibrium prices, p_1^* and p_2^*. These prices are obtained by simultaneously solving the first-order conditions:

$$(1)\qquad \pi'_{Di} = (p_i - v'_i)\cdot Q^i_i + Q^i(p_i, p_j) = 0, \qquad i, j = 1,2,$$

where $v'_i(Q_i, c_i)$ is Di's marginal cost of producing Q_i units of output when it is paying c_i for the relevant input. Q^i_i is the derivative of demand for the product produced by Di with respect to its own price.

The FOCs in equation (1), also generate the best-response functions, $B_i(p_j; c_i)$. From equation (1), comparative statics yield,

$$(2)\qquad \partial B_i/\partial c_i = \left[(\partial v'_i/\partial c_i)\cdot Q^i_i\right]/\pi''_{Di} > 0,$$

as long as the relevant input is normal, as we shall assume from now on. Hence, as the cost of the relevant input to the firm increases, the firm will increase its optimal price, holding rival's price constant.

We assume that the prices of the two final products are strategic complements, in the terminology of Jeremy Bulow et al. (1985), meaning that $\partial B_i/\partial p_j > 0$. This assumption is satisfied if the second cross-partial derivative, Q_{ij}, is not "too negative." It is satisfied, for example, when downstream demands are linear.

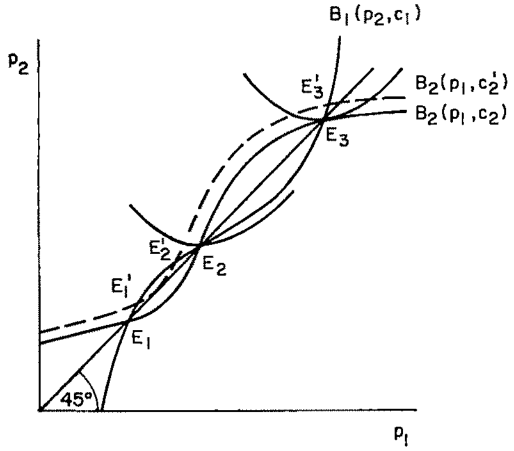


Figure 1. Stable and Unstable Equilibria

These minimal restrictions placed on the demand and cost functions in the final goods market raise the possibility that there may be more than one Bertrand-Nash equilibrium in the downstream market. In the case of multiple equilibria, some are "unstable."[10] In Figure 1, E_1, E_2, and E_3 are the intersections of $D1$ and $D2$'s best-response functions. Equilibria E_1 and E_3 are stable, while E_2 is unstable. The dashed curve represents $D2$'s best-response function when its costs are increased slightly to c_2', leading to new equilibria E_1', E_2', and E_3', respectively. As shown, the movement from E_3 to E_3' increases $D1$'s profits, while the movement from E_2 to E_2' decreases them. (The concave curves through E_2 and E_3 represent $D1$'s isoprofit curves through those points: "higher" isoprofit curves represent higher levels of profit.) These results are summarized as:

Remark 1: An increase in the rival's input cost increases (resp., decreases) the equilibrium prices around a stable (resp., unstable) Bertrand-Nash equilibrium.

Furthermore, since

$$\partial \pi_{Di}/\partial c_j = \left(\partial \pi_{Di}/\partial p_j\right)\cdot\left(\partial p_j^*/\partial c_j\right),$$

then we also have:

Remark 2: An increase in the rival's input cost raises (resp., lowers) the firm's profits around the stable (resp., unstable) Bertrand-Nash equilibrium.

Two other questions regarding the profit effects of cost shifts remain: (i) whether or not an increase in the firm's own input cost may plausibly increase its profits; and (ii) whether or not an industrywide increase in costs increases a firm's profits. Now, an increase in the firm's own costs causes a change in profits given by:

$$(3)\qquad \partial \pi_{Di}/\partial c_i = -x_i + (p_i - v_i') \cdot \left(Q_j^i\right)\cdot\left(\partial p_j^*/\partial c_i\right),$$

where x_i stands for Di's use of the input. The sign of this expression is ambiguous around a stable equilibrium. The case of interest is where that derivative is negative so that no firm would like to see its costs raised unilaterally. An industrywide cost increase is more likely to raise firms' profits than a unilateral increase because it leads to a greater change in equilibrium prices. In the case of linear demand (worked out in the Appendix), however, both types of cost increase lower firms' profits. Then, $D1$ and $D2$ will not tacitly collude to impose higher costs upon themselves.

D. *The Standard Foreclosure Argument*

With these preliminaries out of the way, we can state the foreclosure argument as follows. $D1$ acquires $U1$, say, and prevents it from selling the input to the unintegrated rival $D2$. $U1$'s withdrawal from the market potentially increases $U2$'s market power over the price of the input. In the market setting posited here, $U2$ becomes a monopolist supplier of the input to $D2$ and sets the price at level c_2^m which maximizes its profits:

---

[10] See Jesus Seade (1980), Avinash Dixit (1986), or Jean Tirole (1988, pp. 323 ff.) for discussions of stability in oligopoly models. Stability is related to the convergence of dynamic adjustment processes. In this context, one imagines the thought experiment in which firm 1 chooses a price, then firm 2 chooses its static best-response to that price, then firm 1 chooses its best-response to firm 2's best-response, etc. If this process converges to the equilibrium, the equilibrium is said to be stable. Of course, in a static context, there is no practical content to the thought experiment.

$$(4)\quad \pi_{U2}(c_2) = (c_2 - m)\cdot x_2^*(c_2; c_1 = m),$$

where m is the marginal cost of the input to $U1$. Obviously, $c_2^m > m$, which is the level that obtains under Bertrand competition in the input market.[11]

Focusing on stable equilibria, since $D1$'s profits are increasing in its rival's costs, the strategy of backward vertical integration makes $D1$ better off and lowers rival's profits. In addition, the strategy also increases downstream prices of both products. Because there are no efficiency gains from vertical integration, social welfare falls. Hence, on all accounts, $D1$'s integration with $U1$ is anticompetitive.

The preceding derivation of the anticompetitive consequences of the vertical merger is, however, woefully inadequate. First, why should $U1$ withdraw from the merchant market after the merger? May it not be in the interests of the merged $D1$-$U1$ firm to (partly) constrain $U2$'s market power by standing ready to supply $D2$ at some price? Second, why would $D2$ not be able to employ some counterstrategy that would protect its profit margins, for example by merging with $U2$? Finally, should $D2$ not be allowed to compete with $D1$ for the right to acquire the upstream supplier in the first place? And would such competition not render $D1$'s initial anticompetitive strategy unprofitable?

In the following section we consider a richer model which not only sheds some light on these questions but allows foreclosure through vertical merger to emerge as a fully rational, equilibrium strategy in a well-specified game.

II. A Game-Theoretic Model of Vertical Foreclosure

A. *Description of the Game*

In order to incorporate all the features of interest, we analyze a four-stage game. In the first stage, the "bidding" stage, the downstream firms have an initial opportunity to acquire one of the upstream suppliers. If there is only one merger, following the convention of the previous section, we suppose it to be between $D1$ and $U1$ and denote the merged firm by $F1$.

In the second stage, input prices are determined.[12] As we shall see, how input prices are set is a crucial determinant of the equilibrium of the overall game. It is therefore important to be careful how one models the second stage.

If no vertical merger takes place in stage one, $D1$ and $D2$ pay marginal cost for the input. If there is an acquisition, $D1$ continues to obtain the input at its true marginal cost. However, it is less clear what assumption about the price $D2$ must pay is most reasonable. We consider and contrast three possibilities:

(i) $F1$ and $U2$ simultaneously choose the input prices at which they are prepared to supply $D2$ (a la Bertrand).

(ii) $F1$ can commit to not supplying $D2$ at all. In this case only we consider the possibility that there exists an alternative, albeit inferior, source of the relevant input. The input can be obtained from that source for $\underline{c} > m$. This assumption enables us to parametrize the upstream price. Note, in particular, that if $\underline{c}$ is sufficiently high, $U2$ has monopoly power over $D2$, as in the previous section.

(iii) $F1$ is able to set an upper bound on the price that $D2$ faces by standing ready to supply it at a price c_{12}, which is determined by $F1$. To capture this, we consider the case in which $F1$ announces its linear price to $D2$ *before* $U2$ announces its price.

The major modeling difference is between (ii) and (iii) on one hand and (i) on the other. When (i) pertains, the merged firm is unable to resist the temptation of competing with $U2$ for $D2$'s business. However, a farsighted $F1$ will realize that entering into competition of this kind will lower $D2$'s input price, which, in turn, will lead to lower downstream prices. The downstream division of $F1$ would, generally, prefer that the upstream division exercises some pricing restraint. Assumption (ii) captures this restraint. However, it may not be in $F1$'s interest to completely withdraw from the upstream market. Indeed, as we shall show, it may not be in the interest of firm 1 to endow $U2$ with "too much" market power over to price of the input. Assumption (iii) provides $F1$ with the ability to constrain $U2$'s market power in the post-merger game.

---

[11] This is true as long as the two downstream products are not perfect substitutes.

[12] Since input prices are determined before the output prices, it is proper to characterize the input suppliers as the upstream firms, as one of the referees reminds us.

In sum, as one moves from (i) to (iii), $F1$ has increasing control over the price of the input to $D2$. We show how the equilibrium of the game depends on this assumed degree of control.

In the third stage, if $D1$ acquired $U1$, $D2$ can bid for its supplier, $U2$. This allows us to analyze $D2$'s ability to counter the potentially adverse consequences of the previous merger by becoming vertically integrated itself. Since organization formation is less flexible than choices of prices, one might think that both opportunities for vertical mergers should logically come before the selection of input prices. Our focus, however, is on the desirability of the second merger assuming that the first has taken place. Considerations about the rationality of the second merger depend on the input price that $D2$ believes it will face if it does not merge. Thus considering the setting of input prices in stage 2 is a way of providing the relevant backdrop against which $D2$ makes its merger decision.

In the fourth stage, downstream prices are chosen, given input prices.

B. *Solving for the Equilibrium*

To solve for the equilibrium, we proceed by backward induction.

(i) *Stage Four*. This stage is simply the one-shot game analyzed in Section I.

(ii) *Stage Three*. In this stage, we inquire whether $D2$ should make a bid for $U2$, and whether it can make a bid that is attractive to $U2$. The answer to this question depends on the sum of the profits earned by $U2$ and $D2$, if they were to remain independent, as compared to those they would earn after a merger. If the latter profits are higher than the former, then the $U2$-$D2$ merger generates gains and we would expect such gains to be realized in an efficient bargain and the merger to take place.

If $D2$ and $U2$ merge, there will be full vertical integration in the industry. The profits earned by the merged entities will be $\pi_{Di}(m,m) \equiv \pi_i^O$, $i=1,2$, which are the levels that are realized absent integration and with Bertrand competition in the input market. We refer to these profits as "benchmark" profits. If $D2$ and $U2$ do not merge, $U2$ charges $D2$ an input price c_2 (which is determined in the second stage) and the sum of their profits, which we term "stand-alone profit" and denote by π^S, is

$$(5) \quad \pi^S(c_2) = \pi_{D2}(c_2) + \pi_{U2}(c_2, m).$$

$D2$ and $U2$ will not merge if and only if $\pi_{U2}(m,m) < \pi^S(c_2)$.

We must show that there exists some $c_2 > m$ for which this inequality is satisfied. To this end we differentiate equation (5) with respect to c_2. After using the envelope theorem and Shephard's lemma, we get:

$$(6) \quad \partial\pi^S/\partial c_2 = (c_2 - m)\cdot(dx_2/dc_2) + (\partial\pi_{D2}/\partial p_1)\cdot(\partial p_1^*/\partial c_2),$$

where dx/dc denotes the total equilibrium change in $D2$'s demand for the input following its increase in price. Note that, around the stable equilibrium, the first term is negative and the second is positive. Thus, when (6) is evaluated around $c_2 = m$ it leads to:

PROPOSITION 1: *If the initial equilibrium is stable, then a small increase in the price of an input above marginal cost to the unintegrated downstream firm increases the "stand-alone profit," π^S, defined in equation* (5).

The import of Proposition 1 is that, for at least some range of the input price c_2, the stand-alone profit of the two unmerged firms increases above the benchmark level, π_{D2}^0. This means that $D2$ cannot make an attractive enough merger offer to its (monopolistic) upstream supplier, $U2$. Consequently, if

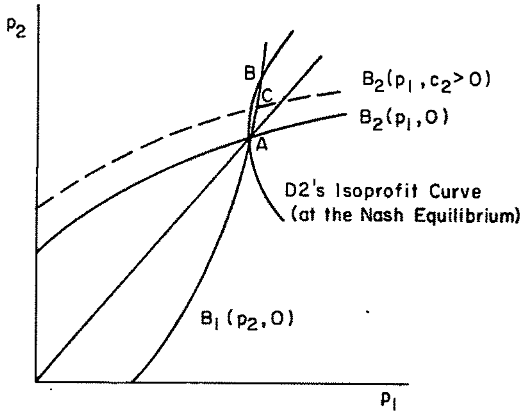


FIGURE 2. BEST-RESPONSE FUNCTIONS WITH PRICE AS THE STRATEGIC VARIABLE

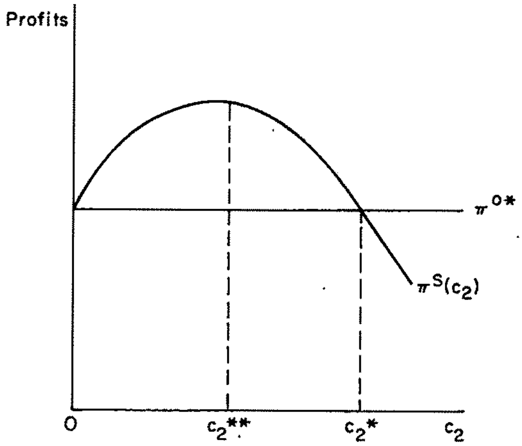


FIGURE 3. THE SUM OF PROFITS OF THE FORECLOSED FIRM AND ITS SUPPLIER

the equilibrium value of c_2 is in that range, $U2$ cannot resort to a merger as an effective counterstrategy.

The reason that such a range exists in the model analyzed here can be seen from Figure 2 which illustrates the benchmark equilibrium. There the locus of points which give $D2$ its benchmark level of profits is superimposed on the best-response curves. Since $D2$ has no costs and $U2$ has no revenues in the benchmark case, that curve also represents the locus of points where the sum of $D2$ and $U2$'s revenues are equal to the benchmark level.

A small increase in c_2 above m causes $D2$'s best-response function to shift up (as represented by the dashed line), and therefore p_1^* and p_2^* to increase, which increases downstream revenues (the new equilibrium, C, lies on a higher iso-revenue curve for $D2$). While $D2$ must pay more for the input, since this additional payment is captured by $U2$, combined profits are equal to $D2$'s revenues which increase.

Eventually a further increase in the input price decreases the "stand-alone" profit. As the markup $(c_2 - m)$ grows, the equilibrium output of $D2$ falls. Ultimately, $D2$'s revenues fall; then the profits earned by $U2$ are insufficient to compensate $D2$ for its lost profits. This occurs in Figure 2 when $D2$'s best-response function intersects $D1$'s best-response function above B. Stand-alone profit cannot fall below the benchmark level, π_{U2}^0, because that level can always be secured via merger. Figure 3 depicts the behavior of the stand-alone profit as a function of c_2.

(iii) *Stage Two*. In this stage input prices are determined. The only substantive case to consider is when $D1$ and $U1$ have merged to form $F1$. We have to consider three alternative pricing scenarios. In the first, $F1$ and $U2$ simultaneously choose the price to be charged to $D2$. Let c_{12} denote the price that $F1$ charges $D2$. Plainly for any c_2, $F1$ has an incentive to undercut, in the usual Bertrand fashion. So, in equilibrium $c_{12} = c_2 = m$.

The second alternative is that $F1$ commits not to sell to $D2$. Consider first the case where $D2$ does not have access to an inferior input substitute. Regardless of whether a $U2$-$D2$ merger subsequently occurs $U2$ earns $\pi_{U2}(c_2, m)$. This is because if a merger occurs, $D2$ must offer a high enough price to compensate $U2$ for what it would earn if it did not merge. Thus $U2$ sets the monopoly price c_2^m. If, however, $D2$ does have access to a substitute inferior product priced at $c < c_2^m$, then $D2$ will set c_{12} just below c.

The third, and most interesting alternative pricing scenario is when $F1$ offers to supply $D2$ at some price c_{12}. $U2$'s best reply to any such $c_{12} > m$ is to undercut it infinitesimally. Given this opportunity to constrain $U2$'s market power by precommiting to c_{12}, $F1$

selects that price that makes the sum of stand-alone profits, π^S, equal to benchmark profits, π^0_{U2}, as calculated in Section I. In Figure 3, this input price is denoted by c_2^*. When input is priced at that level, $D2$ cannot make a merger offer that $U2$ would not refuse.

For the linear-linear case examined in the Appendix, $c_2^* < c_2^m$ so that $F1$ does exercise a competitive constraint over $U2$'s pricing power vis-à-vis $D2$. It is not clear whether this holds for all relevant demand and cost configurations.

To summarize the second stage: If $U2$ and $F1$ move simultaneously, then $c_2 = m$. If $F1$ commits not to supply $D2$ then $c_2^* = c_2^m$ in the case where there is no alternative source of supply and is equal to the lesser of c_2^m and c when there is one. Finally, if $F1$ can set a price to constrain $U2$, then $c_2 = c_2^*$.

(iv) *Stage One.* In this stage the initial acquisition(s) can be made. There are numerous ways in which one could formulate the actions in the acquisition stage. Bidding could be initiated either by the buyers or the sellers of the input. A bidder could bid for only one firm at the time or it could make multiple offers. Alternatively, each firm could be paired with a single target. Here we examine the simplest case where both firms bid for a designated upstream firm, say $U1$. The robustness of the results is taken up in Section III.

To see how this stage plays out, we first compare the payoffs from winning versus losing this initial bidding contest. For consistency, we assume that $D1$ is the winning bidder, in the event that there is a bid. Recall that there will be no bid if both firms correctly expect $D2$ to acquire $U2$ when it has a chance to deploy its counterstrategy in stage 3. Such acquisition will take place if the input is overpriced in stage 2. Then, with both firms vertically integrated profits are again at the "benchmark" level.[13]

So, the only interesting case is when the input is priced at c_2^*, so that $D2$ will be just indifferent between buying and not buying its supplier. How much is $D1$ willing to bid for $U1$? Note that if there is a merger, $D1$, as an owner of $F1$, makes $\pi_{D1}(p_1^*(m, c_2^*), p_2^*(c_2^*, m); m)$ on the downstream account. If $D1$ submits a losing bid, it will earn $\pi_{D1}(p_1^*(c_1^*, m), p_2^*(m, c_1^*); c_1^*)$. Here c_1^* is the input price that $D1$ will have to pay in the event that $D2$ prevails in the bidding game. Obviously, the maximum amount that $D1$ will be willing to bid, W^1, is the difference between these two profit levels. By an earlier assumption, downstream firm's profits decline as its costs increase unilaterally. Hence, $W^1 > 0$. In the symmetric case $W^1 = W^2 = W$. Given the assumed structure of the remainder of the game, $U1$ earns zero if it does not accept the offer so it accepts $D1$'s bid of W.[14]

C. *Equilibrium Summarized*

The unique equilibrium of the game is as follows:

If $F1$ and $U2$ set input prices simultaneously, so that $c_2 = 0$; or if $F1$ can only commit not to supply $D2$ without being able to constrain $U2$ so that $U2$ wants to charge $c_2^m > c_2^*$ and there is no substitute input at $c < c_2^*$, then no vertical merger takes place.

If $F1$ can commit not to supply $D2$ and there is a substitute input at $c < c_2^*$, then both firms bid $W(c)$ for $U1$. The offer is accepted and $D1$ faces the input cost of m while $D2$ pays c.

If there is no substitute product and $F1$ can commit to supply $D2$, both firms bid $W(c_2^*)$. $U1$ accepts the offer and $D1$ faces the input cost of m while $D2$ pays c_2^*.

D. *Welfare in Equilibrium*

It is immediate that a vertical merger between $U1$ and $D1$ raises both downstream prices. Obviously, consumers' welfare falls.

---

[13] Notice that we do not allow Di to make a bid for a Ui in Stage 3, in the event that the initial merger did not take place.

[14] Note that $D2$ cannot induce $U1$ to remain independent by offering $U1$ a lump-sum payment, L. Plainly, by pitting $D1$ against $D2$, $U1$ can always earn W. Hence L would have to be greater than W. But $D2$ is not willing to pay more than W.

TABLE 1—THE INCENTIVE TO INTEGRATE IN THE LINEAR DEMAND CASE

d (1)	$p_i^*(0,0)$ (2)	π_i^O (3)	c_2^* (4)	$p_1(0,c_2^*)$ (5)	$p_2(c_2^*,0)$ (6)	$\pi_{D1}(0,c_2^*)$ (7)	$\frac{\pi_{D1}(0,c_2^*)}{\pi_i^O}$ (8)	$\frac{p_1(0,c_2^*)}{p_1^*(0,0)}$ (9)
0	0.50	0.25	0.00	0.500	0.500	0.250	1.000	1.000
1	0.33	0.22	0.09	0.345	0.381	0.238	1.073	1.036
2	0.25	0.19	0.13	0.274	0.321	0.225	1.200	1.095
3	0.20	0.16	0.14	0.229	0.278	0.210	1.315	1.147
5	0.14	0.14	0.13	0.175	0.219	0.183	1.492	1.222
10	0.08	0.08	0.09	0.110	0.142	0.133	1.743	1.320
$d \to \infty$	0.00	0.00	0.00	0.000	0.000	0.000	2.250	1.500

The unintegrated downstream firm, $D2$, is made worse off as a result of the merger because only its input cost went up while the rival's cost remained at m. More surprisingly, perhaps, former $D1$ is also worse off: its net payoff is the same as $D2$'s in this symmetric game. Its gross profits are increased by the merger, of course. However, because it paid W to acquire $U1$ its net profits are lower than they would have been had there been no vertical merger: both downstream firms would be better off if the bidding war was never started—if publicly traded, their stock prices fall when the bidding commences! The game has a prisoners' dilemma structure: any firm that believes that the other will not bid for $U1$ can gain by buying up $U1$ at a low price and capturing the rents from the foreclosure of its rival. The fear of being foreclosed drives each firm to attempt to foreclose the other. As a result, all the rents from foreclosure are dissipated through the bidding and all the profits accrue to the upstream firm(s). In a richer model, with asymmetric downstream payoffs, for example, the foreclosing firm also could be made better off in equilibrium.

The clear winners thus are the upstream firms. $U1$ gets $W(c_2^*)$, while $U2$ earns $\pi_{U2}(c_2^*)$. Note that $U1$ is better off than $U2$. This follows from the fact that π_1^* exceeds π^S, due to the inefficient pricing of the input to $D2$. However, $D1$ and $D2$ earn the same *net* profits. Thus, $U1$'s shareholders obtain a better return on their shares than do $U2$'s shareholders. This is an attractive feature of the model because it obviates the possibility of the upstream target holding out hoping that the other firm will be taken over.

The net effect of a merger is a reduction in social welfare. This is because the merger does not remedy any existing market imperfection but, instead, leads to an inefficient pricing of the input to $D2$. The fact that additional profits accrue to the input suppliers does not overcome the losses to consumers.

While our analysis illustrates that vertical foreclosure may be rational, with the degree of generality in the above model, it is not possible to measure the size of the incentives to integrate there. In the Appendix we explicitly calculate the equilibria with and without a vertical merger for the case of linear demand for the case where the integrated firm is able to constrain the price which $U2$ can charge if there is a $U1$-$D1$ merger (scenario (iii)). We assume there that demand for good i is given by $q_i = 1 - bp_i + d(p_j - p_i)$, and that $m = 0$. We have carried out some calculations, setting $b = 1$.

Table 1 reports the results of those calculations for different values of d, which measures the sensitivity of each firm's demand to the difference in the firms' prices.

Notice from columns (2) and (6) that as d grows, competition between the downstream firms becomes so fierce that prices are driven to marginal cost, irrespective of whether there has been a merger or not. Therefore (as one can see from columns (3) and (7)), firm 1's profits vanish in the limit too.

Columns (8) and (9) present the results of most interest. The increase in $D1$'s profitability is given in (8). Clearly if $d = 0$ $D1$'s profitability is the same whether it merges with $U1$ or not. As d increases, however, the merger becomes increasingly profitable. In the limit, as $d \to \infty$, the merged firm's profits are increased by 125 percent.

Because profits become insignificant as d increases, it is possible that even a tiny percentage increase in price can result in a large percentage increase in profitability. In column (9), therefore, we directly examine the effect of a merger on $D1$'s equilibrium market price. While the percentage increases in price are smaller than those in profits, the figures are still sizable.

III. Robustness

In this section, we discuss the robustness of the results of Section II to three changes in the assumptions. First, we examine equilibrium outcomes when downstream firms compete in quantities rather than prices. Second, we discuss changes in the bidding formulation: we discuss what happens if the downstream firms can make offers to either of the upstream firms in stage 1, or if the upstream firms bid for a downstream firm. Finally, we discuss more general vertical contracts, such as two-part tariffs.

A. *Quantity Competition*

We now consider what happens when the downstream firms' actions are strategic substitutes. To do this we assume quantity competition in stage four, that the products are perfect substitutes, and that the unique Cournot equilibrium is stable. Under reasonable assumptions about marginal costs and revenues, Di's best reply to an expansion of output by Dj is to constrict its output.

Using the approach of Section II, we calculate the effect of an increase in c_2 on π^S, the stand-alone profits of $U2$ and $D2$. Differentiation gives:

$$(8)\quad \partial\pi^S/\partial c_2 = (c_2 - m)\,dx_2/dc_2 + (\partial\pi_{U2}/dq_1)\cdot(\partial q_1/\partial c_2).$$

Around $c_2 = m$, the sign of this expression is governed by the sign of the second term. The sign of that term is unambiguously negative because an increase in c_2 induces expansion in $D1$'s output which, in turn, causes p_2 to fall for any given level of q_2. This depresses $D2$'s profits. Hence:

PROPOSITION 2: *When the downstream firms compete in quantities, the only equilibrium of the four-stage game entails no vertical merger.*

The intuition here can be obtained by reference to the taxonomy of strategic actions developed by Drew Fudenberg and Jean Tirole (1984). An increase in $D2$'s costs makes it "soft" (i.e., less aggressive) in its competition with $D1$. Accordingly in the case of strategic complements, apart from the fact that it must bear the higher costs itself, it favors a "fat-cat" posture. That is, if it has a less aggressive position vis-à-vis $D1$, more accommodating pricing results. In the case of strategic substitutes, however, $D2$ wants to maintain a "lean-and-hungry" look. That is, any gain in output and profitability for the $U2$-$D2$ "side of the market" must be gained at the expense of the $U1$-$D1$ side.

Thus the $U2$-$D2$ side of the market is made strictly worse off in aggregate if the price charged by $U2$ rises. Accordingly, $U2$ and $D2$ will always counter a $U1$-$D1$ merger with a merger of their own.

The relevant test of whether a vertical merger for the purposes of foreclosure can be successful, therefore, is whether the sum of the profits of the unintegrated firms will increase or not.

B. *Alternative Bidding Formulations*

As was discussed above, there are many possible alternative formulations of the bidding stages. Here we briefly discuss just two.[15]

First, suppose that the downstream firms can make offers to either upstream firm. If we continue to assume that the downstream firms simultaneously make take-it-or-leave-it offers to the upstream firms, then the equilibrium analyzed in Section II remains an equilibrium. If *U*2 believes that *U*1 will accept *D*1's offer, then there does not exist an offer that *D*2 can profitably make which will be acceptable to *U*2.[16]

---

[15] These cases are discussed in more detail in our working paper, Janusz Ordover, Garth Saloner, and Steven Salop (1988).

Second, suppose that it is the upstream firms that bid for a downstream firm rather than *vice versa*. Thus, consider the case where in stage 1, *U*1 and *U*2 bid for *D*1 and in stage 3, *D*2 can become the target if the first merger occurred.

The main effect from this reformulation of the game is the reallocation of rents and the size of the winning bid in stage 1. The post-merger input and products' prices will be exactly as they would have been had it been the downstream firms bidding for the upstream supplier.[17]

C. *Alternative Vertical Contracts*

We now discuss alternatives to integration as ways of raising downstream prices. Our discussion focuses on the case of differentiated products and price competition, in light of Proposition 2.

To begin with let us note that where it is possible for firms to enter into alternative contractual relationships with their input suppliers, these may appear to be preferable to a merger. The purpose of these alternatives, like that of a fully-fledged vertical merger, is to raise downstream prices by affecting input costs.[18]

The key point of this section is easily illustrated with the help of Figure 3. There, with $c_1 = m$, c_2^{**} maximizes π^S, the sum of *U*2 and *D*2 profits. In order to induce *U*2 to charge that price rather than c_2^*, *D*2 can enter into a two-part tariff contract with its supplier under which the fixed fee transfer compensates *U*2 for not charging c_2^*. (This point is made by Bonanno and Vickers, 1988.) In our setting vertical integration is just the special form of two-part tariff where the price is equal to marginal cost and the fee is the purchase price.

Two observations can now be made. First, even if *U*2-*D*2 counter the *U*1-*D*1 merger with an optimal two-part tariff, the initial merger remains profitable. However, the bid price will be lower because the equilibrium downstream prices will be lower.

Second, it may appear that *U*2-*D*2 could replicate the outcome under a two-part tariff by first merging and then setting the *transfer price* at c_2^{**}. This conclusion is implausible because once these two firms have merged, there is an incentive to cut the transfer price below c_2^{**}. Furthermore, the rival who cannot easily monitor intra-firm pricing will, most likely, proceed on the expectation that transfer prices are set at efficient levels. As a result, the superior equilibrium cannot be replicated with a merger because a firm cannot credibly commit to internal transfer prices that are inefficient, in our view.[19]

---

[16] Of course there are other similar equilibria as well: *U*2 could be designated as the target and either *D*1 or *D*2 could be selected as the winning firms. Notice that providing the downstream firms with all the bargaining power creates an asymmetry in the resulting outcome. Although *U*1 and *U*2 are both viable targets and ought to be expected to compete for the right to be taken over, instead the downstream firms "designate" the winner who then gets the lion's share of the spoils. Formally, since the integrated firms do better than the unintegrated firms, requiring the upstream firms to each do equally well and the downstream firms to each do equally well would have to be accomplished by utilizing mixed strategies (with some resulting inefficiency) or resorting to correlated equilibria.

[17] Another effect of this reformulation is that a holdout problem (discussed in detail in our working paper) can arise. To see how this might arise, consider the products are almost independent in demand. There the upstream firm that merges gains very little because the profits that accrue to the merged firm are only slightly increased by the increase in the input price to the unmerged rival. The unmerged upstream firm, in contrast, is suddenly endowed with monopoly power over the remaining downstream firm. Thus both upstream firms would prefer to be the firm that does not merge.

[18] This point is similar to that made by Cooper (1986), who points out that in a differentiated products context a firm (and its rival) may be made better off if it can commit to charging a higher price than the noncooperative equilibrium price. In Cooper's model the firm achieves this by offering a "most-favored-customer" contract which, once it has been implemented, gives the firm an incentive not to lower its price in the future.

The distinction between two-part tariffs and vertical integration can be brought out again by considering a game in which in the first stage, $D1$ and $D2$ present $U1$ with two-part tariff contracts. For example $D1$ may bid $\{f = W(c_2^*) > 0,\ c_1 = 0,\ c_2 = c_2^*\}$ which potentially replicates the equilibrium outcome under vertical integration. Notice, however, that the vertical contract must specify the price which $U1$ must charge $D2$. Moreover, since $U2$ has an incentive to secretly cheat on this contract and supply $D2$ at a lower price, the restrictive contract will only be feasible if $D1$ can monitor $U1$'s price. However, a contract that specifies $U1$'s merchant market prices is likely to run afoul of the antitrust laws. In contrast, a vertically integrated firm can more credibly commit to its merchant market price and may refuse to sell the input to the rival without risking antitrust litigation.[20]

IV. Conclusions

We have investigated how vertical integration can be used to achieve anticompetitive foreclosure of downstream firms from an essential source of supply. Our main conclusion is that anticompetitive foreclosure arises as an equilibrium phenomenon in a coherent model where sophisticated firms use a wide range of strategies and counterstrategies.

Our analysis indicates that the profitability of foreclosure involves a number of subtleties. However these subtleties are not the ones associated with the standard critique of foreclosure theory discussed in the introduction. Indeed, the central condition for successful foreclosure is simply that the unintegrated upstream firm's gain exceed the downstream firm's loss.

This condition can be satisfied in a model with differentiated products and price-setting because the sum of the profits of the foreclosed firm and its supplier increase (locally) if the foreclosed firm's price rises (and the rival's price adjusts optimally). In a quantity-setting, homogeneous goods model, by contrast, the sum of those profits decrease if the foreclosed firm's output decreases and the rival's adjusts optimally.

Moreover, even in the differentiated products case, the sum of the unintegrated firms' profits is reduced if the unintegrated upstream firm's market power is "too" great. This is because, in that case, the profit-maximizing price that it charges for its input will be so high that the downstream firm will have the incentive and ability to offer a mutually beneficial vertical merger. In this paper we have set out two of the ways in which the unintegrated upstream firm's market power may be limited. One is that the integrated firm may be able to constrain the upstream spot price. The other is that there may be an alternative (and inferior) competitive source of supply.

As another alternative, one could model the upstream industry as an oligopoly rather than a duopoly. In that case, a single vertical merger would not confer complete monopoly power on the remaining upstream firms. While competition among them would naturally constrain the upstream price, reducing the profitability of foreclosure, such competition also would reduce the potential for an effective counterstrategy on the part of the foreclosed downstream firm. The necessary modeling is not entirely straightforward, however. Since the interesting case is one in which the elimination of an upstream firm does have some effect on input prices, this requires specifying an upstream market in which the goods are at least somewhat differentiated (perhaps because of transportation costs). Enriching the model in this way is a possible avenue for future research.

The controversy surrounding the vertical foreclosure argument has traditionally been cast in the context of vertical integration. In order to address the questions raised in that debate we have, for the most part, kept to that formulation. Yet, as modern contract theory emphasizes, vertical integration is but one particular form of vertical contract. Indeed, as we have shown, there may exist incentives for firms to structure vertical relationships in ways that fall short of full vertical integration if they are so able. This richer formulation raises a number of interesting issues that we have only begun to explore.

---

[19] Chaim Fershtman and Kenneth Judd, (1987) for example, analyze a model in which firms distort managers' compensation schedules in order to change their conduct in the product market. As here, there are strong incentives to cheat on these inefficient arrangements, especially in light of the fact that such internal arrangements are not easily monitored for deviations.

[20] It may also be easier to commit to a two-part tariff than an internal transfer price by coupling the two-part tariff with a most favored nation clause as in Salop (1986).

APPENDIX

Here we solve the model in the "linear-linear" case. Thus, we assume the following: (i) upstream unit costs are equal to zero, $m=0$; (ii) downstream marginal cost is equal to c_i, $i=1,2$; and the demand functions are symmetric, that is, $q_i = 1 - bp_i + d(p_j - p_i)$, $i, j = 1,2$. It is useful to rewrite this as $q_i = 1 - kp_i + dp_j$, where $k = b + d$.

Our main objective is to analyze how the equilibrium prices, input cost, and purchase price vary with the value of d, which determines the degree of substitution between the two products. These comparative statics are reported in the text in Table 1.

Standard calculations yield the following reduced form equations:

$$p_i^* = \left[(2k+d) + 2k^2c_i + kdc_j\right] \Big/ (4k^2 - d^2), \tag{A1}$$

$$q_i^* = k\left[(2k+d) - (2k^2 - d^2)c_i + kdc_j\right] \Big/ (4k^2 - d^2), \tag{A2}$$

$$\pi_{Di}^* = (q_i^*)^2/k. \tag{A3}$$

When input is priced at $c_1 = c_2 = 0$, Di earns

$$\pi_{Di}^0 = k/(2k-d)^2. \tag{A4}$$

When $c_1 = 0$, $U2$'s profits are:

$$\pi_{U2}(c_2) = kc_2\left[(2k+d) - c_2(2k^2 - d^2)\right] \Big/ (4k^2 - d^2). \tag{A5}$$

Hence, $U2$'s unconstrained profit maximizing price is:

$$c_2^m = (2k+d)/2(2k^2 - d^2). \tag{A6}$$

The maximum input price that forestalls the $U2$-$D2$ merger, denoted by c_2^*, solves

$$\pi_{U2}(c_2) + \pi_{D2}^*(c_2, c_1 = 0) = \pi_{D2}^0. \tag{A7}$$

Thus,

$$c_2^* = (2k+d)\,d^2/2k^2(2k^2 - d^2) < c_2^m, \tag{A8}$$

since $b, d > 0$.

REFERENCES

Aghion, Philippe and Bolton, Patrick, "Contracts as a Barrier to Entry," *American Economic Review*, June 1987, *77*, 388–401.

Bonanno, Giacomo and Vickers, John, "Vertical Separation," *Journal of Industrial Economics*, March 1988, *36*, 257–66.

Bork, Robert H., *The Antitrust Paradox*, New York: Basic Books, 1978.

Bulow, Jeremy I., Geanakoplos, John D., and Klemperer, Paul D., "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, June 1985, *93*, 488–511.

Comanor, William S., "Vertical Mergers, Market Power, and the Antitrust Laws," *American Economic Review*, May 1967, *57*, 254–65.

______ **and Frech, H. E.,** "The Competitive Effects of Vertical Agreements," *American Economic Review*, June 1985, *75*, 539–46.

Cooper, Thomas E., "Most-Favored-Customer Pricing and Tacit Collusion," *Rand Journal of Economics*, Autumn 1986, *17*, 377–88.

Coughlin, Ann and Wernerfelt, Birger, "Competition and Cooperation in Marketing Channel Choice: Theory and Application," *Marketing Science*, Spring 1988, *4*, 110–29.

Dixit, Avinash, "Vertical Integration in Monopolistically Competitive Industry," *International Journal of Industrial Organization*, September 1983, *1*, 63–78.

Fershtman, Chaim and Judd, Kenneth, "Equilibrium Incentives in Oligopoly," *American Economic Review*, December 1987, *77*, 927–40.

Fudenberg, Drew and Tirole, Jean, "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean-and-Hungry-Look," *American Economic Review*, May 1984, *74*, 361–66.

Krattenmaker, Thomas G. and Salop, Steven C. "Anticompetitive Exclusion: Raising Rivals' Costs to Achieve Power Over Price," *Yale Law Journal*, December 1986, *96*, 209–93.

Mathewson, G. Frank and Winter, Ralph, "The Competitive Effects of Vertical Agreements: Comment," *American Economic Review*, December 1987, *77*, 1057–62.

Moorthy, K. Sridhar, "Strategic Decentralization in Channels," *Marketing Science*, forthcoming.

Perry, Martin K., "Vertical Integration: Determinants and Effects," in Richard Schmalensee and Robert Willig, eds., *The Handbook of Industrial Organization*, New York: North-Holland, 1989, 185–255.

Posner, Richard A., *Antitrust Law*, Chicago: University of Chicago Press, 1976.

______ **and Easterbrook, Frank H.,** *Antitrust*, 2nd ed., St. Paul: West Publishing Co., 1981.

Rey, Patrick and Stiglitz, Joseph, "The Role of Exclusive Territories in Producers' Competition," Princeton University, mimeo., September 1986.

______ **and Tirole, Jean,** "The Logic of Vertical Restraints," *American Economic Review*, December 1986, *76*, 921–39.

Salinger, Michael A., "Vertical Mergers and Market Foreclosure," *Quarterly Journal of Economics*, May 1988, *77*, 345–56.

Salop, Steven C., "Practices that (Credibly) Facilitate Oligopoly Coordination," in Joseph Stiglitz and C. Frank Mathewson, eds., *New Developments in the Analysis of Market Structure*, Cambridge: MIT Press, 1986, 265–90.

______ **and Scheffman, David T.,** "Cost-Raising Strategies," *Journal of Industrial Economics*, 1987, *36*, 19–34.

Schwartz, Marius, "The Competitive Effects of Vertical Agreements: Comment," *American Economic Review*, December 1987, *77*, 1063–8.

Seade, Jesus K., "On the Effects of Entry," *Econometrica*, March 1980, *48*, 479–89.

Tirole, Jean, *The Theory of Industrial Organization*, Cambridge: MIT Press, 1988.

Disadvantageous Oil Tariffs and Dynamic Consistency

By ERIC MASKIN AND DAVID NEWBERY*

A large importer who places relatively greater weight on future than current oil consumption will import less oil in the future than if it were able to commit itself in advance to future tariffs, and may find itself worse off than if it were unable to impose tariffs at all. Futures markets and storage modify these adverse effects and may avoid the problem of dynamic inconsistency. (JEL 026)

The analysis of natural resource taxation has largely concentrated on the taxation of resources extracted within the legal jurisdiction of the government.[1] With few exceptions,[2] the analysis of import tariffs on natural resources has been neglected, presumably on the grounds that it raises no issues not covered in the standard international trade literature.

Theodore Bergstrom (1982) is one of these exceptions, and he argues that tariffs on oil have a special and attractive property. He observes that excise taxes on inelastically supplied goods fall entirely on the supplier. If oil is available in a fixed amount and is (essentially) costless to extract, then its total supply will be inelastic. If it is competitively supplied, then a constant *ad valorem* tax will fall entirely on the supplier, and in a closed economy the government could use the tax to extract the entire rent. With international trade in oil and many noncooperative consuming nations, he shows that if each nation chooses the optimal constant *ad valorem* tariff, taking the choices of other governments as given, and if suppliers are competitive, then again the entire incidence of the tariffs will fall on producers and cause no distortion.[3]

Bergstrom's paper is mainly concerned with the exercise of monopsony power on the part of importers, and it thus falls within the standard optimal tariff framework. Two other arguments for tariffs have been advanced, though neither will be considered here. The first is the argument that if supplies of oil are subject to random interruptions then it may be optimal to impose a tariff to reflect the increased vulnerability that oil imports imply. (See the discussion in William Nordhaus, 1974, and, for a more recent treatment of policies to reduce vulnerability, James Plummer 1982.) The other argument is that of countervailing power against the OPEC cartel.[4] In practical policy

---

*Department of Economics, Harvard University, Cambridge, MA 02138 and Department of Applied Economics, Cambridge University, Cambridge, England CB3 9DE. This paper is an extensively revised version of Maskin and Newbery (1978). We thank the NSF, the Sloan Foundation, and the British Economic and Social Research Council for research support, and Ted Bergstrom for his excellent comments.

[1] See, for example, Mason Gaffney (1967); Gerard Brannon (1975); Partha Dasgupta, Geoffrey Heal, and Joseph Stiglitz (1980); Ross Garnaut and Anthony Clunies Ross (1983).

[2] Since our (1978) working paper pointing out the likely dynamic inconsistency of an open loop import tariff, several other papers have appeared that explore consistency issues in the international oil market. As we explain below, these papers work with quite different models and for the most part ask different questions.

[3] This argument continues to go through if the marginal cost of extraction is positive but constant. It fails, however, if marginal cost is an increasing function of the quantity extracted. Indeed, Larry Karp (1984) shows that a monopsonist importer's optimal open-loop tariff will be dynamically inconsistent if competitive sellers face increasing extraction costs. If the importer could impose a fixed tax in addition to the *ad valorem* tariff on any seller supplying a positive quantity of oil, then it could extract all the rent from the oil and restore the dynamic consistency of the open-loop tariff even in the presence of increasing costs. Neither conclusion remains if, as in our model below, there are competitive buyers as well as the large importer.

[4] See Theodore Bergstrom, John G. Cross, and Richard Porter (1981) for the case of bilateral monopoly.

discussions it is clearly necessary to recognize the bilateral monopoly (or oligopoly) aspect of the world oil market, but for a theoretical understanding it seems to us worthwhile to clarify a problem that arises with tariffs even with monopsony power alone.[5]

There is a significant difference between the choice of an optimal tariff on an ordinary (perishable)[6] produced good, and that on an imported exhaustible resource such as oil: in choosing an optimal tariff on a normal good, all the country need announce is the current rate of the tariff, since whether the supplier wishes to supply now depends only on the current price. In the case of oil, suppliers will only sell now if to do so is more advantageous than keeping the oil in the ground for later extraction and sale. They thus need to compare current oil prices with predicted future prices, and these will in turn depend on future levels of the tariff. In principle, then, when choosing an oil import tariff, the government should be choosing a time path for the oil import tariff. (Thus, in Bergstrom's 1982 model the government chooses a constant *ad valorem* rate, which producers expect to continue into the future, and to which they respond.)

If governments can commit themselves to an announced time path of import tariffs, (i.e., to an "open-loop" strategy), then standard optimization techniques can be used to find the optimal path of the tariff, which, as will be shown below, has a particularly simple and intuitive form. The critical issue is, however, whether governments can so commit themselves. Even within a national jurisdiction there are obvious difficulties facing a government wishing to commit its successors, and there is certainly widespread skepticism on the part of oil producers about the stability of announced tax regimes. But even if governments could overcome this local difficulty, they would find it more difficult to do so on the international front, for it is the essence of national sovereignty to be free to change trading arrangements when it is advantageous. International agreements such as the GATT may have some effect in providing stability in trade tax regimes, but it is notoriously easy to circumvent that agreement by apparently domestic taxes, or by other means. When it comes to trade in such a strategically sensitive product such as oil, it is equally clear that neither long-term contracts on the part of suppliers nor long-term commitments on the part of buyers are normally credible—neither can be readily enforced since no supranational body has the power to impose effective sanctions on those who break contracts.

In this paper we explore the implications of the inability of countries to commit to future tariffs. We show precisely why the optimal open-loop tariff (the optimal tariff assuming commitment) is dynamically inconsistent, that is, why an importer would wish to change the time path of the tariff in midstream. More importantly, we shall point out the qualitative differences between equilibrium with and without commitment.

The literature on the dynamic inconsistencies of open-loop policies is, by now, quite large. It embraces macroeconomic policy (Finn Kydland and Edward Prescott, 1977), the exploitation of a free access resource

---

Bergstrom (1982) briefly discusses the case in which all importing countries choose the *same* constant *ad valorem* tariff against a single monopoly supplier, but this will be optimal only in very special circumstances. Albert Nichols and Richard Zeckhauser (1977) consider the related problem of choosing the optimal stockpile to countervail a monopoly supplier, which, as David Newbery (1984) shows, can be done better by the choice of a countervailing tariff. All these papers suffer from the difficulty that the monopolist is modeled as implausibly passive—see Newbery (1984).

[5] Jonathan Eaton and Zvi Eckstein (1984) study several models with both a large importer and exporter (although they also include a brief discussion of the 1978 version of this paper). Rather than analyze how the inability of parties to commit themselves to future actions affects the time paths of prices and quantities (i.e., computing the open-loop and dynamically consistent paths), they concentrate on the role that oil reserves (storage) can play in improving the importer's position. However, they run into the usual difficulty of modeling timing that bilateral monopoly creates: does the importer or exporter move first or do they move simultaneously?

[6] Goods that can be stored share some of the durability properties of exhaustible resources, and hence face similar problems to those discussed here. See Stokey (1982), and the discussion of storage later.

(Jennifer Reinganum and Nancy Stokey, 1985, and David Levhari and Leonard Mirman, 1980), and sales by a durable good monopolist (Jeremy Bulow, 1982, and Stokey, 1982), among other topics. There is, moreover, a small set of papers that deals with inconsistency in exhaustible resource settings. Alistair Ulph and Folie (1980) and David Newbery (1981) (which builds on Eric Maskin and David Newbery, 1978) treat the case of monopoly sellers, whereas, in addition to the papers mentioned above,[7] Murray Kemp and Long (1980),[8] and Eckstein and Martin Eichenbaum (1985)[9] examine large importers. Indeed, there is an interesting contrast between modeling large suppliers and large buyers in the absence of commitment. In the former case, the open-loop strategy is dynamically inconsistent only in certain cases, and for a large class of other cases will be dynamically consistent. (See Newbery, 1981.) As we shall see below, the open-loop optimal tariff is always dynamically inconsistent except in one borderline case.

In Section I, we present the model and exhibit the conventional open-loop optimal tariff for a large importer. In Section II we argue that this tariff policy is almost certain to be dynamically inconsistent. We also define the dynamically consistent equilibrium that would prevail if the importer could not commit to future tariffs. In a two-period (present/future) version of the model, we show that this latter equilibrium entails lower oil prices, lower future and total imports, and higher tariffs than in the open-loop equilibrium, if the large oil importer places greater relative weight on future imports (compared to current imports)—as would be the case for an importer with rapidly depleting oil reserves, like the United States.[10] Just the reverse is true if the importer places comparatively *less* weight on future imports.

The inability to commit itself to future tariffs can serve only to lower the importer's welfare. In Section III we show by example that if it places sufficiently greater weight on the future, the large importer may actually be made worse off for having the ability to impose tariffs than if it had no tariff setting power at all. That is, it may be harmed by its monopsony power.

The analysis through Section III assumes that futures markets in oil are absent and that the importing country cannot store oil cost effectively. In Section IV we argue that the availability of a futures market can overcome the dynamic inconsistency of the open-loop tariff. If, however, spot markets continue to operate in the future, the importer may actually have to sell oil, only to buy it back later. We also show that the ability to store oil can serve as a partial substitute for a futures market.

I. The Open-Loop Tariff

Suppose that the marginal cost of extracting oil is c. Suppose also that at date T a cheap alternative to oil (for example, fusion) will become available in unlimited quantities at a constant marginal cost $c^* < c$. If oil producers act competitively, that is, as price takers, and face prices p_t at date t, then each will choose an extraction/sales rate s_t to maximize the present discounted value of profits, subject to the condition that cumulative extraction does not exceed the total initial stock, S. Because $c^* < c$, no oil producer will be able to operate profitably after date

---

[7] That is, those by Karp (1984) and Eaton and Eckstein (1984), which, as noted above, use quite different models and emphasize different issues from those treated here.

[8] Murray Kemp and Ngo Long (1980)—like David Newbery (1976) (the immediate precursor to the present paper) on which it builds—recognizes the inconsistency of the open-loop tariff policy but fails to exhibit the consistent tariff.

[9] Zvi Eckstein and Martin Eichenbaum (1985) do not explicitly model sellers. Instead, they simply postulate an inverse supply function with no assurance that this derives from a seller optimization problem. Their model, therefore, is not suitable for considering the strategic interaction between buyers and sellers that is the concern of this paper.

[10] Strictly speaking, this is true only if the importer has no discretion over the rate of domestic depletion, otherwise the problem is more like that of optimal storage, considered in Section IV.

T. The producer's problem, therefore, is to choose s_t to maximize

$$\sum_{t=1}^{T} \{(p_t - c)s_t\}\delta^{t-1}, \tag{1}$$

subject to

$$\sum_{t=1}^{T} s_t \le S, \tag{2}$$

where δ is the discount factor. Hence, in any period in which s_t is positive,

$$(p_t - c)\delta^{t-1} = \lambda, \tag{3}$$

where λ is the Lagrange multiplier for (2). This gives the standard arbitrage condition

$$p_t - c = \delta(p_{t+1} - c). \tag{4}$$

Henceforth, assume that the marginal utility of oil at zero consumption is larger than c. Then, in equilibrium, s_t is positive in every period and the initial price determines the entire trajectory. This boundary condition is determined by market clearing:

$$\sum_{t=1}^{T} d_t = \sum_{t=1}^{T} s_t,$$

where d_t is demand in period t.

Now consider the problem of choosing the amount of oil to import, x_t, for a country large enough to influence the world market price of oil. Suppose that the demand for oil by all other consumers in aggregate is $y_t(p)$.[11] Suppose that the importer announces a time path x_t at date zero, and all producers believe that the plan will be followed. There is full information about $y(p)$, stocks, and extraction costs, so that all producers can predict future prices and choose their extraction and sales plans at date zero. We assume that neither is there a futures market in oil, nor can consumers store oil for future consumption (but see Section IV where these assumptions are dropped). The importer's problem is to choose x_t to maximize domestic welfare, where the dollar benefit of oil at date t is $U(x, t)$. The discounted value of benefits less costs is given by

$$\sum_{t=1}^{T} \{U(x_t, t) - p_t x_t\}\delta^{t-1}, \tag{5}$$

and this has to be maximized subject to the resource constraint

$$\sum_{t=1}^{T} \{x_t + y_t(p_t)\} \le S, \tag{6}$$

and subject to the arbitrage condition (4) that producers are willing to supply at each date. If we let the domestic price of oil be $q_t = \partial U(x_t, t)/\partial x$, then $q_t - p_t$ represents the importer's tariff on oil. The optimal tariff satisfies

$$\tau_t \equiv q_t - p_t = \tau_1\delta^{t-1}, \tag{7}$$

where τ_1, the initial tariff, is just the Lagrange multiplier for (6). Notice that if the other consumers were not present so that $y_t(p) = 0$, the importer could extract all the producers' profit by adjusting the tariff so that $p_1 = c$. This reflects the general proposition that an import tariff rising at the rate of interest acts as a lump sum tax since oil is in fixed supply.

To call the solution, $x^*(\cdot)$, of the program (4)–(6) the "importer's optimum" presumes that the importer can commit itself to this time path.[12] If it cannot make such a commitment, then at time $\hat{t} > 0$ it will face the problem

$$\text{Max} \sum_{t=\hat{t}}^{T} \{U(x_t, t) - p_t x_t\}\delta^{t-\hat{t}}, \tag{8}$$

---

[11] Hence the other consumers are assumed to be price takers and their current demand is independent of future prices. This can be derived from preferences which are additively separable in the oil consumption of different periods and linear in other commodities.

[12] We shall sometimes argue as though the importing country's policy instrument is its level of imports. Setting an import level, of course, is equivalent to the more conventional policy of setting a tariff.

subject to (4) and

$$\sum_{t=\hat{t}}^{T}\{x_t+y_t(p_t)\}\le S_{\hat{t}},$$

where $S_{\hat{t}}$ is the stock remaining at time $\hat{t}$. Let $\hat{x}(\cdot)$ be the solution to the latter program. As we shall see in the next section, only by mere happenstance will $\hat{x}_t$ and x_t^* coincide for $t\ge\hat{t}$. But without coincidence, x_t^* cannot be regarded as the importer's optimum, because, once period $\hat{t}$ arrives, the importer will prefer an alternative time path.

Clearly, in a world where it cannot predetermine its future oil imports (or, equivalently, cannot bind itself to future import tariffs), the importer must optimize subject to the additional constraint that it should not later wish to change its plan. After all, announced future import levels that will later be changed have no bearing on rational suppliers, who can forecast what the real levels will be. In the remainder of this paper, we explore the nature of the importer's optimum under this additional constraint.

II. Dynamically Consistent Tariffs

When the importer cannot commit itself in advance to a tariff trajectory, we must consider the relative timing within each period: does the importer set its tariff before, simultaneously with, or after the sellers choose their supplies? We wish to argue that it makes the most sense to suppose that the importer moves first.[13] Suppose that we enrich the model slightly to allow for small random shocks each period to the importer's oil requirements (i.e., to its utility function). Assume, furthermore, that the realizations of these shocks are better known to the importer than to the sellers. Then, because in any period the import tariff will depend on the realization of the shocks, the sellers are better off waiting until the importer has set its tariff before determining their supplies. This would be true even if a seller could, at a cost, learn the value of the realization directly: given its small size, the seller cannot affect the importer's behavior by moving earlier, and so therefore can delay its selling decision without cost (i.e., it will not gain from learning the value and so will not pay the cost). Note that this reasoning does *not* symmetrically apply if the sellers, as well as the importer, are subject to shocks. Because it is big, the importer profits from moving first and so would be willing to sacrifice the additional information it could acquire by delay.

We can solve for the importer's optimum by dynamic programming. Suppose that in the final period T the remaining stock of oil is S_T. The importer chooses the level of imports, x, knowing that this will affect the import price, p, through the market clearing relations $x+y(p)\le S_T$. The dependency of price on imports is best handled as a standard constrained maximization problem,

$$\max_{x,p}\{U(x,T)-px\},\tag{9}$$

subject to

$$x+y_T(p)\le S_T.$$

Let $x_T(S_T)$ and $p_T(S_T)$ be the maximizing choices of x and p, and define

$$V_T(S_T)\equiv U\{x_T(S_T),T\}-p_T(S_T)x_T(S_T).$$

Proceeding by backward induction, the importer faces the following problem in period t if the remaining stock is S_t:

$$\max_{x,p}\left[U(x,t)-px+\delta V_{t+1}\{S_t-x-y_t(p)\}\right]\tag{10}$$

subject to

$$x+y_t(p)\le S_t$$

---

[13] We are grateful to Oliver Hart for discussions on this point.

and the arbitrage condition

$$p-c=\delta\left[p_{t+1}\{S_t-x-y_t(p)\}-c\right].$$

Let $x_t(S_t)$ and $p_t(S_t)$ be the maximizing values of the control variables, and take $V_t(S_t)$ to be the maximized value of the objective function. Solving these T iterative programming problems determines the dynamically consistent equilibrium.

To study the qualitative properties of this equilibrium, we shall simplify by making the following assumptions:

A1 There are two periods ($T=2$), 'now' and 'the future,' and the marginal extraction cost c is zero.

A2 The preferences of the large importer are $U(x,1)=aU_1(x)$, $U(x,2)=U_2(x)$, where the U_i are increasing and strictly concave ($U_i'>0$, $U_i''<0$), and a is a parameter to be varied.

A3 The rest of the world has demand $y_i(p_i)$ in period i, where

$$y_i'<0,\quad p_iy_i''(p_i)+2y_i'(p_i)<0$$
$$<p_iy_i''(p_i)+y_i'(p_i),$$

and

$$\frac{y_2''(p)}{y_2'(p)}\geq\frac{\delta y_1''(\delta p)}{y_1'(\delta p)}.$$

These conditions are sufficient to ensure that the first-order conditions guarantee an optimum, as shown in the Appendix, and are satisfied for $y(p)=b_ip^{-\varepsilon}$, $0<\varepsilon\leq1$—a constant elasticity of demand schedule. The large importer's preferences can be written

$$(11)\quad aU_1(x_1)+\delta U_2(x_2)-p_1x_1-\delta x_2p_2.$$

For numerical illustrations (see Section III and the Appendix) we consider the example $U_1(x)=\delta\log(x)$, $U_2=\log(x)$, $y_1(p)=\delta b/p$, $y_2(p)=1/p$. In this case, the importer's objective function is

$$(12)\qquad \delta a\log(x_1)+\delta\log(x_2)$$
$$-p_1x_1-\delta p_2x_2.$$

In the Appendix we give the analytical solution to this example. (These functional forms may obscure the fact that here—as in the rest of the paper—the discount factor δ is common to everyone.)

Suppose that there exists a_0 such that if $a=a_0$ the open loop equilibrium coincides with the dynamically consistent equilibrium. (In the example the critical value is $a_0=b$, at which point all importers value future consumption equally relative to current consumption.) We claim that at all values $a\neq a_0$, the open loop equilibrium is not dynamically consistent. To see why this is so, it will be helpful to consider how the importer would behave in the second period if it had the opportunity to renege on its tariff commitment. The following proposition is our main result.

PROPOSITION 1: *Given assumptions (A1–A3) above, suppose there exists a_0 such that if $a=a_0$ the open loop (binding contract) equilibrium (x_i^*, p_i^*, τ_i^*) coincides with the dynamically consistent equilibrium (x_i^e, p_i^e, τ_i^e). Then if $a>a_0$,*

$$(i)\quad p_2^{**}>p_2^*,\quad x_2^{**}>x_2^*,\quad \tau_2^{**}<\tau_2^*,$$

and

$$(ii)\quad p_i^e>p_i^*,\quad x_2^e>x_2^*,\quad \tau_2^e<\tau_2^*,$$
$$x_1^e+x_2^e>x_1^*+x_2^*,\quad i=1,2,$$

where $(x_2^{**}, p_2^{**}, \tau_2^{**})$ are the large importer's optimal choices in the second period if the binding contract is reneged. The inequalities go the other way if $a<a_0$.

The proof is given in the Appendix. It is not difficult to explain this result intuitively. If the importer places comparatively greater weight on first-period oil consumption, (i.e., $a>a_0$), it would like to ensure that a relatively low first-period price prevails. But it must also induce the competitive producers to supply at this low price rather than holding off until the second period. In other words, it is constrained by the arbitrage condition (4). The importer satisfies the arbitrage condition—that is, it also keeps the second price low—by committing itself to a high second-period tariff (thereby restraining second-period demand). Once the second period actually arrives, however, there is no longer a need for a high tariff; the arbitrage condition is now irrelevant because the first-period oil has already been sold. Thus if the importer can somehow get out of the commitment, it will reduce the tariff, import more oil, and thereby drive the market price up. The oil suppliers will also profit from this reneging, although, of course, if they had anticipated its occurrence, they would not have been willing to sell in the first period.

If $a < a_0$, the argument is just the reverse. The solution now requires an exceptionally *low* tariff in the second period in order to keep the market price high enough to induce producers to sell then. With the arrival of the second period, the importer will have the incentive to raise the tariff, and, if successful, will make the suppliers regret that they did not sell all their oil in the first period.[14]

This sort of argument also explains the discrepancy between the open loop and dynamically consistent equilibria. The proposition implies that for $a > a_0$, $p_i^e > p_i^*$, $i = 1,2$, $x_2^e > x_2^*$, and $\tau_2^e < \tau_2^*$. To see why these relations hold, recall that the second-period reneging price p_2^{**} is greater than the open loop price p_2^*. Then, if $p_2(S_2)$ is the optimal second-period price for the large importer when the stocks at the start of period 2 are S_2 (the solution to (6) in the Appendix) $p_2^{**} = p_2(S_2^*)$, and so

$$p_1^* < \delta p_2(S_2^*). \tag{13}$$

To satisfy the arbitrage equation in the dynamically consistent equilibrium, the importer must bring the left- and right-hand sides of (13) into equality, that is, it must (a) make first-period prices more attractive and (b) make second-period prices less attractive to suppliers. But (a) implies that $p_1^e > p_1^*$, and, since $p_2^e = p_1^e/\delta$, it must be that $p_i^e > p_i^*$, $i = 1,2$. Hence $y_i^e < y_i^*$, $i = 1,2$.

To accomplish (b), and drive down the second-period price, S_2 must be made larger than S_2^*. Since $p_2^e > p_2^*$ and $y_2^e < y_2^*$, $S_2^e > S_2^*$ implies that $x_2^e > x_2^*$. Finally, $(x_2^e, p_2^e) > (x_2^*, p_2^*)$ implies that $\tau_2^e > \tau_2^*$.

Without making stronger assumptions, we cannot say which of x_1^e and x_1^* is larger when $a > a_0$. Although $p_1^e > p_1^*$ and $S_2^e > S_2^*$ (which suggests that x_1^e should be less than x_1^*), it is possible for $p_1^e + \tau_1^e$ to be less than $p_1^* + \tau_1^*$ (which would imply that $x_1^e > x_1^*$). (In the example given in the Appendix it is possible to show that $x_1^e < x_1^*$.)

Proposition 1 therefore gives an almost complete characterization of the dynamically consistent equilibrium relative to the easily solved open loop equilibrium, in which the optimal tariff would rise at the rate of interest—the only incompleteness concerns the level of first-period imports. The proposition starts by defining a benchmark for the relative importance of consumption in the two periods for the large importer. The benchmark is that weight on first-period consumption which provides no incentive to renege on the open loop equilibrium. The main result can be summarized as follows. If the large importer places greater weight on first-period consumption (relative to the benchmark weight), then the dynamically consistent equilibrium will be characterized by a higher price, a higher second-period level of imports and a larger total level of imports by the large importer, than in the open loop equilibrium. The inability to commit to the open loop tariff is advantageous to oil producers but harmful to the large importer and other consumers. In the opposite case in which the large importer places relatively less weight on first-period consumption, prices are lower, other consumers are better off, and the large importer and the suppliers are worse off than in the open loop equilibrium.

---

[14] This may seem to suggest that when $a < a_0$ the importer can implement the binding contract equilibrium, even when commitment is impossible, by using an import quota: the importer can set $x_1 = x_1^*$ as in the open loop equilibrium and refuse to import more than x_1^* in an attempt to keep up the first-period price. Of course, rational suppliers will forecast a fall in next period's price and will try to unload all their oil in the first period. While the import quota stops them selling more to the large importer, the fall in price will induce other importers to increase their demand, and so the price will fall below p_1^* after all.

This suggests that one should ask whether there are simple examples in which the open loop equilibrium is dynamically consistent, for then one can consider perturbations about these special cases. Newbery (1976) shows that if the large importer and the rest of the world each have unchanging demand schedules of constant elasticity, and if extraction costs are zero, then the only case in which the open loop tariff is dynamically consistent is that in which the demand elasticities are the same. The result can be generalized slightly as follows.[15] Suppose that oil is costless to extract and must be used up by date T. At date t demand by the rest of the world is $y(p)=b_t p^{-\varepsilon}$, and by the large importer is $x_t = a_t q^{-\alpha}$, where q is the tariff-inclusive price, and a_t and b_t vary over time. At date t the optimal open loop tariff is given by

$$\tau = \frac{X_t}{\varepsilon Y_t} = \frac{p_t^{\varepsilon-\alpha}(1+\tau)^{-\alpha}}{\varepsilon} \times \frac{\int_t^T a_u e^{-\alpha r u}\,du}{\int_t^T b_u e^{-\varepsilon r u}\,du},$$

where X_t and Y_t are the remaining amounts of oil which will be consumed between t and T by the large importer and the rest of the world, with $S_t = X_t + Y_t$ being remaining stock. (The derivation is given in the Appendix.) Thus if $\varepsilon=\alpha$, and $a_t = kb_t$, then the *ad valorem* tariff rate is constant (as required for the open loop tariff), and there is no temptation to deviate. The share of imports by the large importer remains the same and so the future always looks the same. If $\alpha=\varepsilon$, but a grows more slowly than b (that is, the large importer places relatively more weight on the present), then the formula shows that the importer would like to lower the tariff in the future and import more than in the open loop equilibrium. The case will be just as in Proposition 1 with $a > a_0$.

III. Disadvantageous Monopsony

The inability to commit to future tariffs clearly acts as a constraint on the importer's behavior. The importer's welfare is lower in the dynamically consistent equilibrium than in the open loop equilibrium (since, with commitment, the importer always has the option of reproducing the dynamically consistent quantities). What we shall argue now is that, in the absence of commitment, the importer's market power may actually work to its *disadvantage*. That is, its very ability to set import tariffs may leave it worse off than had it had no tariff-setting powers at all.[16]

To make this point we must first describe the equilibrium that would emerge were the importer unable to set tariffs. We shall refer to this as the *competitive* equilibrium. In the two-period model, a competitive equilibrium is a vector of prices and imports $(p_1^c, p_2^c, x_1^c, x_2^c)$ such that

$$\frac{\partial}{\partial x}\{U(x_t^c,t) - p_t^c x_t^c\} = 0, \qquad t=1,2, \tag{14}$$

$$\sum_{t=1}^{2}\{x_t^c + y_t(p_t^c)\} = S, \tag{15}$$

and

$$p_1^c = \delta p_2^c, \tag{16}$$

where, as before, other importers choose $y_t(p_t^c)$ and (16) reflects the assumption that extraction costs are zero.

Because there are no tariffs in the competitive equilibrium, prices must be higher than in either the open loop or dynamically consistent equilibria. We first show that, in the

---

[15] The suggestion to pursue this line of inquiry comes from Bergstrom.

[16] We should point out that, although we use the term "disadvantageous monopsony," the phenomenon we are concerned with is completely distinct from that considered by Robert Aumann (1973).

TABLE 1—COMPARISON OF EQUILIBRIA

Variable	Values in Various Equilibria Competitive	Open-Loop	Consistent	Reneged
p_1	7.2	7.025	6.788	7.025
p_2	7.2	7.025	6.788	6.131
x_1	0.028	0.024	0.025	0.024
x_2	0.139	0.122	0.091	0.101
W	1.109	1.125	1.079	1.173
τ_1/p_1	0.0	0.171	0.177	0.171
τ_2/p_2	0.0	0.171	0.618	0.618
S_2	0.278	0.264	0.238	0.264

case $a > a_0$, prices in the dynamically consistent equilibrium must be intermediate between the competitive and open loop prices. To see this, notice that, because it is always desirable to have a positive tariff,

$$p_2^c > p_2(S_2^c), \tag{17}$$

where again $p_2(S_2)$ is the optimum second-period price for the large importer given stock S_2. Thus, p_1^c and $p_2(S_2^c)$ do not satisfy the arbitrage condition. To bring them into line, we must reduce p_1 (and raise $p_2(S_2^c)$). Hence the dynamically consistent first-period price, p_1^e must be less than p_1^c, and so $p_1^* < p_1^e < p_1^c$ by the Proposition.

The welfare of the large importer is a concave function of prices, with a maximum at the open loop prices, and so when $a > a_0$, the importer's welfare in the dynamically consistent equilibrium is between the welfare levels for the competitive and open loop equilibria.

By contrast, when the importer places greater relative weight on second-period imports than do other consumers ($a < a_0$) the dynamically consistent prices are *less* than the open loop prices. Thus we can no longer bound the importer's welfare between the competitive and open loop levels. Indeed, as the following example shows, if the importer places sufficient relative weight on the second period, its dynamically consistent welfare may actually be less than the competitive level.

Example: Consider the logarithmic utility function and unit elastic demand schedules of the example of equation (12) (analyzed in the Appendix), and let $a = 0.2$, $b = 5$, $S = 1$, and $\delta = 1$. We obtain Table 1, where W refers to the total importer welfare (plus 5), and τ_i/p_i $(i = 1,2)$ is the proportional tariff in period i.

In this example, the importer's incentive to raise the tariff in the second period of the open loop equilibrium is so strong that, for dynamic consistency, it is forced to set an exceptionally high second-period tariff and, consequently, to import exceptionally little oil. Given the importance of second-period oil to welfare, this ability to set high tariffs turns out to be the importer's undoing.

IV. Futures Markets and Storage

In Sections I–III we assumed that futures markets do not exist and that consumers cannot store oil. This meant that all consumption arose directly from spot purchases. We next investigate how our results change in the presence of forward trading or storage.

Clearly, if a futures market for date 2 oil merely replaced the date 2 spot market, then the importer could duplicate its open loop program in our two-period model, and the issue of commitment to future tariffs would not arise. All the importer would have to do at date 1 is to buy x_1^* on the spot market (at price p_1^*) and x_2^* on the forward market (at price $\delta p_2^* = p_1^*$).[17]

---

[17] We speak of the importer as doing the purchasing, as though it were government policy. But of course no government intervention is needed: the consumers buy x_i^* at price $p_i^* + \tau_i^*$.

The more delicate question arises when there is a futures market in *addition* to the date 2 spot market. In that case, it will not suffice for the importer simply to buy x_1^* and x_2^* at date 1 because there is nothing preventing it from reentering the market at date 2 and altering its date 2 consumption.

To see the issues involved, consider the importer's problem at date 2, assuming that it purchased z units of oil on the futures market[18] and that S_2 is the remaining stock (including z). The importer chooses x_2^* and p_2 to

$$\text{(18)} \qquad \max_{p_2, x_2} U_2\{x_2\} - p_2(x_2 - z),$$

subject to

$$x_2 + y_2(p_2) \le S_2.$$

The first-order conditions are

$$\text{(19)} \qquad U_2' - p_2 - \lambda = 0,$$

$$\text{(20)} \qquad -(x_2 - z) - \lambda y_2' = 0,$$

and

$$\text{(21)} \qquad x_2 + y_2(p_2) - S_2 = 0.$$

Suppose that the importer tried to implement the open loop equilibrium quantities. If we replace the variables in (19)–(21) with their open loop values, the left-hand sides of those equations become

$$\text{(22)} \qquad U_2'(x_2^*) - p_2^* - \lambda^*,$$

$$\text{(23)} \qquad -(x_2^* - z) - \lambda^* y_2'(p_2^*),$$

and

$$\text{(24)} \qquad x_2^* + y_2(p_2^*) - S_2^*.$$

By definition of the open loop equilibrium, (22) and (24) equal zero. Moreover, we can always choose a value of z (possibly negative) so that (23) is zero too. Thus the importer can, after all, ensure the open loop outcome, even though it cannot prevent the spot market opening in the second period.

Now, if $a < a_0$, we know from Proposition 1 that $x_2^{**} < x_2^*$, $p_2^{**} < p_2^*$, so $y_2'(p_2^{**}) < y_2'(p_2^*)$. Consequently, at $z = 0$, (23) is negative, where (23) is the marginal gain from raising p_2 above p_2^*. Hence, in this case the optimal z is positive. However, if $a > a_0$, the optimal z is negative, that is, it is optimal for the importer to *sell* oil forward and buy it back in the second period. This peculiar behavior is simply a device for ensuring that the importer has no incentive to raise the price of oil in the second period: the more oil it has to buy, the lower the market price it prefers.

Turning to storage, we see that if we reinterpret z as the amount of oil the importer buys and stores, then (19)–(21) are the importer's first-order conditions in the second period, assuming that storage is possible and costless but that futures markets are absent. Of course, negative values for z now make no sense. Thus we conclude that in the case $a < a_0$, the possibility of storage enables the importer to implement the open loop equilibrium,[19] whereas, in the case $a > a_0$, storage makes no difference at all to the dynamically consistent equilibrium (since any worthwhile deviation from zero storage would require negative, infeasible, levels of z).

Although storage is of no use in the case $a > a_0$, it has a dramatic effect on the possibility of disadvantageous monopsony, since, from Section IV, this is possible only if $a < a_0$. In this case storage, if feasible and not too costly, restores the open loop option and eliminates the disadvantage. A strategic oil reserve may thus make sense for a large importing country faced with rapidly declining domestic oil reserves.

---

[18] Again, we can think of the consumers themselves doing the buying, as long as we interpret the tariff as a subsidy when z is negative.

[19] This result breaks down if extraction is costly (even if storage is not) since, given discounting, it would be preferable to extract the stored oil in the future—it would be better to store the oil in the ground, not above it.

V. Concluding Comments

In this paper we have explored the consequences for a large oil importer of an inability to commit its future import tariff levels. Our main objective has been to derive qualitative results within a general two-period model. We have shown that in comparing the dynamically consistent equilibrium with the committed, or open loop, equilibrium, the critical determinant is the relationship between the rate at which the large importer trades off current versus future oil in consumption. If the large importer places relatively greater weight on current oil consumption (compared to future consumption)[20] it will set a lower tariff on imported oil, and, in the future will import less oil than if it were able to commit itself in advance to future tariff levels (i.e., to an open loop tariff). When the large importer places relatively greater weight on future consumption, the results reverse, and in that case the importer can find itself worse off than were it unable to impose tariffs at all. Only in the knife-edge intermediate case will the open loop tariff be dynamically consistent.

We also examine the extent to which futures markets and the ability to store oil get around the dynamic inconsistency problem. If oil extraction costs and storage costs are both zero, then storage can restore the dynamic consistency of the open loop equilibrium in the case where, otherwise, the importer would reduce the tariff in the future. Thus, in particular, it solves the problem of disadvantageous monopsony. But if either oil is costly to extract (so that its price appreciates less rapidly than the rate of interest), or if oil is costly to store (above the interest costs) then storage can be at best partially successful. Moreover, it is no use at all in the case where the importer is tempted to *raise* the future tariff in the open loop equilibrium.

APPENDIX

Proof of Proposition:
We need to show that if $a > a_0$,

$$p_2^{**} > p_2^*, \quad x_2^{**} > x_2^*, \quad \tau_2^{**} < \tau_2^*, \tag{1}$$

and

$$\begin{gathered} p_i^e > p_i^*, \quad x_2^e > x_2^*, \quad \tau_2^e < \tau_2^*, \\ x_1^e + x_2^e > x_1^* + x_2^*, \quad i=1,2, \end{gathered} \tag{2}$$

where superscript e refers to the dynamically consistent equilibrium, * to the open loop (binding contract) solution, and where $(x_2^{**}, p_2^{**}, \tau_2^{**})$ are the large importer's optimal choices in the second period if the binding contract is reneged. Because the arbitrage condition $p_1 = \delta p_2$ must be satisfied, the binding contract equilibrium solves the program

$$\begin{aligned} \max_{x_1, p_2} \; & aU_1(x_1) + \delta U_2\{S - x_1 - Y(p_2)\} \\ & - \delta\{S - x_1 - Y(p_2)\}p_2 \\ & - \delta p_2 x_1, \end{aligned}$$

where $Y(p_2) = y_1(\delta p_2) + y_2(p_2)$ is total imports over both periods by other countries. The first-order conditions are

$$aU_1' - \delta U_2' = 0, \tag{3}$$

$$-\delta U_2' Y' - \delta(S - x_1 - Y) + \delta p_2 Y' - \delta x_1 = 0, \tag{4}$$

where $Y'(p_2) = \delta y_1' + y_2'$, and $Y'' = \delta^2 y_1'' + y_2''$. Equation (3) gives the time path of the tariff. By definition, $aU_1' = q_1 = p_1 + \tau_1$ and $U_2' = q_2 = p_2 + \tau_2$, and so from (3) and the arbitrage condition, $q_1 = \delta q_2$ and $\tau_1 = \delta \tau_2$. Thus the tariff rises at the rate of interest, confirming the general result of Section III. The matrix of second derivatives is

$$\delta \begin{bmatrix} aU_1''/\delta + U_2'' & U_2''Y' \\ & -U_2'Y'' + (Y')^2 U_2'' \\ Y'U_2'' & +2Y' + p_2 Y'' \end{bmatrix}.$$

Because $U_i'' < 0$ and $p_i y_i'' + 2y_i' < 0$, $y_i'' > 0$, and $y_i' < 0$, the matrix is negative definite. Hence the first-order conditions are sufficient for an optimum.

In the second period, the large importer would like to choose p_2 to maximize

$$\begin{aligned} & U_2\{S - x_1 - y_1 - y_2(p_2)\} \\ & \quad - p_2\{S - x_1 - y_1 - y_2(p_2)\}, \end{aligned} \tag{5}$$

---

[20] As pointed out above, the weight is to be compared with the benchmark weight on current consumption that provides no future incentive to deviate from the open loop equilibrium. The example of the Appendix illustrates that this weight will depend on the consumption pattern of the rest of the world.

that is, to satisfy

$$(6)\qquad -U_2'y_2'+p_2y_2' \\ -(S-x_1-y_1-y_2)=0.$$

Let the solution to this problem be p_2^{**}. Clearly, $p_2^{**} = p_2^*$ at $a=a_0$. Now, just as we showed that the maximand in the binding contract program is concave in p_2, so (5) is concave in p_2. Thus to show that $p_2^{**}>p_2^*$ for $a>a_0$, it suffices to show that, when we replace p_2^{**} by p_2^*, x_1 by x_1^*, and $y_1(\delta p_2^{**})$ by $y_1(\delta p_2^*)$, the left-hand side of (6) is increasing in a at a_0.[21] Differentiating the left-hand side of (6) with respect to a, we obtain

$$(7)\quad \{-U_2'y_2''+Y'(1+y_2'U_2'')+p_2y_2''+y_2'\}\frac{dp_2^*}{da} \\ +\{1+y_2'U_2''\}\frac{dx_1^*}{da}.$$

Implicitly differentiating (3) and (4) with respect to a and solving for dx_1^*/da and dp_2^*/da, we find

$$(8)\quad \frac{dx_1^*}{da}=-\frac{\delta U_1'}{\Delta}\{(p_2-U_2')Y''+2Y' \\ +(Y')^2U_2''\}>0,$$

where $\Delta>0$ is the determinant of the matrix of second derivatives, and

$$(9)\qquad \frac{dp_2^*}{da}=\frac{\delta U_1'}{\Delta}Y'U_2''>0.$$

Formula (7) can be rewritten as

$$(10)\qquad \frac{\delta U_1'}{\Delta}[U_2'\{Y''(1+y_2'U_2'')-y_2''Y'U_2''\} \\ -(2Y'+p_2Y'')(1+y_2'U_2'')] \\ +(p_2y_2''+y_2')\,dp_2^*/da.$$

The last term is positive, and, since $\delta U_1'/\Delta>0$, so is the middle term, while the first term has the sign of

$$U_2'\left\{Y''+\delta y_2'U_2''(-y_1')\left[\frac{y_2''(p)}{y_2'(p)}-\frac{\delta y_1''(\delta p)}{y_1'(\delta p)}\right]\right\}>0,$$

from the conditions on $y_i(p)$. Consequently, (7) is positive, as required. Hence $p_2^{**}>p_2^*$ if $a>a_0$ and, thus, since $y_2^{**}<y_2^*$, $x_2^{**}>x_2^*$ and $\tau_2^{**}<\tau_2^*$.

Let us turn to the dynamically consistent equilibrium. Let $p_2=p_2(S-x_1-y_1)$ be the solution to the program of maximizing (5). (Then $p_2^{**}=p_2(S_2^*)$). Implicitly differentiating (6) with respect to S and solving for $p_2'\equiv\partial p_2/\partial S$, we find

$$(11)\quad p_2'=\frac{1+y_2'U_2''}{-U_2'y_2''+(y_2')U_2''+2y_2'+p_2y_2''}<0,$$

since the numerator is positive and the denominator is negative.

Now, if the big importer chooses x_1 in the first period of the rational expectations equilibrium, the price $\bar{p}_2(x_1)$ must satisfy

$$(12)\qquad \bar{p}_2(x_1)=p[S-x_1-y_1\{\delta\bar{p}_2(x_1)\}].$$

Implicitly differentiating (12) with respect to x_1 and solving for $d\bar{p}_2/dx_1$, we find

$$(13)\qquad \frac{d\bar{p}_2}{dx_1}=-\frac{p_2'}{1+\delta y_1'p_2'}>0,$$

as $p_2'<0$ from (11).

Now, at $a=a_0$, $(x_1^e,p_2^e)=(x_1^*,p_2^*)$. Consider what happens as a rises slightly. We have already seen that x_1^* and p_2^* rise. Suppose, contrary to the claim, that $p_2^e\le p_2^*$—we shall show that this is impossible. First we show that if $p_2^e\le p_2^*$, then $x_1^e<x_1^*$. Suppose not, and that $x_1^e\ge x_1^*$. Since by assumption $p_2^e\le p_2^*$, $y_1^e\ge y_1^*$, and so $S_2^e\le S_2^*$. From (11), $p_2^e=p_2(S_2^e)\ge p_2(S_2^*)=p_2^{**}$. But since $p_2^{**}>p_2^*$, this is inconsistent with $p_2^e\le p_2^*$. This means that $x_1^e<x_1^*$, so that

$$(14)\qquad (x_1^e,p_2^e)<(x_1^*,p_2^*).$$

But, because the matrix of second partials is negative definite, (14) implies that the left-hand sides of (3) and (4) must be positive at (x_1^e,p_2^e). But then a slight increase in x_1^e and p_2^e must be an improvement, and in view of (13), this is possible to do without violating constraint (6). Since this means that the importer was not optimizing, contra hypothesis, this means the original premise was false and so $p_2^e>p_2^*$. From the arbitrage condition, this implies that $p_1^e>p_1^*$. Thus, $y_i^e<y_i^*$, $i=1,2$, and so $x_1^e+x_2^e>x_1^*+x_2^*$. Finally, $p_2^e=p_2(S_2^e)<p_2^{**}=p_2(S_2^*)$, and so by (11), $S_2^e>S_2^*$. Since $p_2^e>p_2^*$, and $y_2^e<y_2^*$, we have $x_2^e>x_2^*$. Finally, $(x_2^e,p_2^e)>(x_2^*,p_2^*)$ implies that $\tau_2^e<\tau_2^*$. □

Example.

Let $U_1(x)=\delta a\log(x)$, $U_2=\log(x)$, so the importer's objective function is given by (12) in the main text. If this is divided through by δ we have

$$(15)\qquad a\log(x_1)+\log(x_2)-p_2(x_1+x_2).$$

Let $y_1(p)=\delta b/p$, $y_2(p)=1/p$, so that $Y(p_2)=(1+b)/p_2$. Then, solving (6) for x_2, p_2:

$$x_2(S_2)=\frac{g-1}{g}S_2;\quad p_2(S_2)=\frac{g}{S_2};$$

---

[21] If (5) is written as $F(p,a)$ (ignoring subscripts), then (6) $F_p=0$ (subscripts now indicating partial derivatives), which gives $F_{pp}\,dp+F_{pa}\,da=0$, or $dp/da=-F_{pa}/F_{pp}$, so $dp/da>0$ if $F_{pa}>0$.

so

$$x_2 = \frac{g-1}{p},$$

where $g = (1+\sqrt{5})/2$. The rational expectations solution is found by substituting these expressions into the objective function

$$\max_{p_1, x_1, S_1} a\log(x_1) + \log\{x_2(S_2)\} - p_2(S_2)\{x_1 + x_2(S_2)\}, \tag{16}$$

subject to

$$p_1 = \delta g/S_2$$

and

$$S_2 = S - x_1 - \frac{\delta b}{p_1} = S - x_1 - b/p_2.$$

These last two conditions together give $x_1 = (S - g - b)/p_2$, and with this the objective function can be written as a function of p_2 alone. The solution is

$$\begin{aligned} x_1^e &= S - G/p_2^e; \\ p_2^e &= \frac{G - 1 + \sqrt{(G+1)^2 + 4aG}}{2S}; \\ x_2^e &= (g-1)/p_2^e, \end{aligned} \tag{17}$$

where $G \equiv g + b$. The open loop equilibrium is found by solving (3) and (4):

$$\begin{aligned} x_1^* &= \frac{a}{1+a}\left\{S - \frac{1+b}{p_2^*}\right\}; \\ x_2^* &= \frac{1}{1+a}\left\{S - \frac{1+b}{p_2^*}\right\}; \\ p_2^* &= \frac{1 + b + \sqrt{(1+b)^2 + 4(1+a)(1+b)}}{2S}. \end{aligned} \tag{18}$$

Comparing (17) and (18), we observe that the two equilibria are the same only in the knife-edge case where $a = b$, that is, where the large importer and other oil buyers place the same relative weights on present versus future consumption.

Constant Elasticity Example.

Suppose that oil is costless to extract and must be used up by date T. At date t demand by the rest of the world is $y(p) = b_t p^{-\varepsilon}$, and by the large importer is $x_t = a_t q^{-\alpha}$, where q is the tariff-inclusive price, and a_t and b_t vary over time. The net welfare of the large importer when $q = p(1+\tau)$ is

$$V(p) = \frac{a_t p_t^{-\alpha}(1+\tau)^{-\alpha}(1+\alpha\tau)}{\alpha - 1}.$$

On the open loop path the optimum *ad valorem* tariff is constant, and the price rises at the rate of interest, r, so the present discounted value of net welfare for the large importer at date t is

$$W_t = \frac{p_t^{-\alpha}(1+\tau)^{-\alpha}(1+\alpha\tau)}{\alpha - 1} \times \int_t^T a_u e^{(1-\alpha)r(u-t)}du. \tag{19}$$

The relationship between p_t and τ is given by the market clearing condition

$$\begin{aligned} S_t &= X_t + Y_t \\ &= [p_t(1+\tau)]^{-\alpha}\int_t^T a_u e^{-\alpha r(u-t)}\,du \\ &\quad + p_t^{-\varepsilon}\int_t^T b_u e^{-\varepsilon r(u-t)}\,du, \end{aligned} \tag{20}$$

where X_t and Y_t are the remaining amounts of oil which will be consumed between t and T by the large importer and the rest of the world, and S_t is the remaining stock. The optimal committed tariff starting at date t is found by maximizing W_t with respect to τ, where from (20),

$$\frac{\partial p_t}{\partial \tau} = \frac{-p_t}{(1+\tau)(1+\varepsilon Y/\alpha X)}.$$

The optimal open loop tariff is therefore given by

$$\tau = \frac{X_t}{\varepsilon Y_t} = \frac{p_t^{\varepsilon-\alpha}(1+\tau)^{-\alpha}}{\varepsilon}\,\frac{\int_t^T a_u e^{-\alpha r u}\,du}{\int_t^T b_u e^{-\varepsilon r u}\,du}.$$

REFERENCES

Aumann, Robert J., "Disadvantageous Monopolies," *Journal of Economic Theory*, February 1973, *6*, 1–11.

Brannon, Gerard, ed., *Studies in Energy Tax Policy*, Cambridge, MA: Ballinger, 1975.

Bergstrom, Theodore C., "On Capturing Oil Rents with a National Excise Tax," *American Economic Review*, March 1982, *72*, 194–201.

______, Cross, John G. and Porter, R. C., "Efficiency Inducing Taxation for a Monopolistically Supplied Depletable Resource," *Journal of Public Economics*, February 1981, *15*, 23–32.

Bulow, Jeremy, "Durable Goods Monopolists," *Journal of Political Economy*, April 1982, *90*, 314–32.

Crawford, Vincent, Sobel, Joel and Takahashi, Ichiro, "Bargaining, Strategic Reserves, and International Trade in Exhaustible Resources," mimeo., 1980.

Dasgupta, Partha, Heal, Geoffrey M. and Stiglitz, Joseph E., "The Taxation of Exhaustible Resources," in Gordon A. Hughes and Geoffrey M. Heal, eds., *Public Policy and the Tax System*, London: Allen & Unwin, 1980, ch. 5.

Eaton, Jonathan and Eckstein, Zvi, in Robert Baldwin and Anne Krueger, eds., *The Structure and Evolution of Recent U.S. Trade Policy*, Chicago: University of Chicago Press, 1984.

Eckstein, Zvi and Eichenbaum, Martin, "Oil Supply Disruptions and the Optimal Tariff in a Dynamic Stochastic Equilibrium Model," in Thomas Sargent, ed., *Energy, Foresight, and Strategy*, Washington: Resources for the Future, 1985.

Gaffney, M. Mason, ed., *Extractive Resources and Taxation*, Milwaukee: University of Wisconsin Press, 1967.

Garnaut, Ross and Ross, Anthony Clunies, *Taxation of Mineral Rents*, Oxford: Clarendon Press, 1983.

Karp, Larry, "Optimality and Consistency in a Differential Game with Non-renewable Resources," *Journal of Economic Dynamics and Control*, October 1984, *8*, 73–98.

Kemp, Murray C. and Long, Ngo V., "Optimal Tariffs and Exhaustible Resources," in M. C. Kemp and N. V. Long, eds., *Exhaustible Resources, Optimality and Trade*, Amsterdam: North-Holland, 1980, Essay 18.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Levhari, David and Mirman, Leonard J., "The Great Fish War: An Example Using a Dynamic Cournot-Nash Solution," *Bell Journal of Economics*, Spring 1980, *11*, 322–34.

Maskin, Eric and Newbery, David M., "Rational Expectations with Market Power: The Paradox of the Disadvantageous Monopolist," Warwick Economic Discussion Paper and Massachusetts Institute of Economics Discussion Paper, Cambridge, MA, 1978.

Newbery, David M., "A Paradox in Tax Theory: Optimal Tariffs on Exhaustible Resources," SEER Technical Paper, Stanford Economics Department, 1976.

______, "Oil Prices, Cartels, and the Problem of Dynamic Inconsistency," *Economic Journal*, September 1981, *91*, 617–46.

______, "Commodity Price Stabilization in Imperfectly Competitive Markets," in George C. Storey, Andrew Schmitz, and Alex H. Sarris, eds., *International Agricultural Trade*, Boulder: Westview Press, 1984, ch. 11.

Nichols, Albert L. and Zeckhauser, Richard L., "Stockpiling Strategies and Cartel Prices," *Bell Journal of Economics*, Spring 1977, *8*, 66–96.

Nordhaus, William D., "The 1974 Report of the President's Council of Economic Advisers: Energy in the Economic Report," *American Economic Review*, September 1974, *84*, 558–66.

Pearce, David W., Siebert, Horst and Walter, Ingo, eds., *Risk and the Political Economy of Resource Development*, London: Macmillan, 1984.

Plummer, James, ed., *Energy Vulnerability*, Cambridge, MA: Ballinger, 1982.

Reinganum, Jennifer F. and Stokey, Nancy L., "Oligopoly Extraction of a Common Property Natural Resource: The Importance of the Period of Commitment in Dynamic Games," *International Economic Review*, February 1985, *26*, 161–73.

Salant, Stephen W., "Exhaustible Resources and Industrial Structure: A Nash-Cournot Approach to the World Oil Market," *Journal of Political Economy*, October 1976, *84*, 1079–94.

Stokey, Nancy, "Rational Expectations and Durable Goods Pricing," *Bell Journal of Economics*, Spring 1982, *12*, 112–28.

Ulph, Alistair M. and Folie, G. Michael, "Dominant Firm Models of Resource Depletion," in David Currie, David Peel, and William Peters, eds., *Microeconomic Analysis*, London: Croom Helm, 1981, ch. 4.

On the Treatment of Anticipated Shocks in Models of Optimal Control with Rational Expectations: An Economic Interpretation

By LEONARDO AUERNHEIMER AND GABRIEL A. LOZADA*

We suggest a general, heuristic method, based on elementary notions of price theory, for dealing with the question of anticipated changes in models of perfect foresight. The method is illustrated for the case of anticipated shocks to the state variable; in this context the standard "shadow price continuity condition" fails, while the method presented here applies. After presenting the general solution, it is used to analyze various examples of preannounced changes. (JEL 020,213)

By modeling economic agents as explicitly maximizing some objective function, and by assuming rational expectations (perfect foresight, in deterministic models), one can treat a class of problems of wide applicability and considerable theoretical interest, namely, the effects of preannounced or "anticipated" changes. Such changes can be, of course, either in behavioral or policy parameters or in the level of endogenous variables which otherwise are given by past history—the typical, textbook case of the latter being the usual "earthquake" experiment which destroys part of the capital stock. In the monetary literature, for example, the classic work by Thomas Sargent and Neil Wallace (1973) is one of the earliest examples of this type of analysis.

In all these cases, the announcement (say, at time $t = 0$) of the event to take place at some later time $t = \tau$ does not change any of the parameters until that time (and not even then in the "earthquake type" experiments, in which the event is a once-and-for-all exogenous change of what otherwise is an endogenous variable). The problem is then to find a path of the original system for which, at the time of the change, the relevant variables can embark on the saddle path toward the final steady state while simultaneously satisfying the condition that no unexploited "capital gains or losses" take place at that time. It is only natural for this condition to be required, as *unexploited* capital gains or losses in such a world can only occur at the time of surprises, and in this case the surprise takes place not at the later date at which the event materializes, but when it is announced.

In most cases the condition that there be no *unexploited* capital gains or losses at time τ implies continuity in the price of assets at that time. This is the "asset price continuity principle," and it renders immediate the solution to problems for which is is applicable, that is, cases in which assets are traded in a market; in the Sargent-Wallace analysis mentioned above, for example, the requirement is that the price level be the same immediately before and immediately after the preannounced change.[1] In cases in which there are no asset markets, and hence no asset prices to calculate capital gains or losses, the continuity principle applies, instead, to shadow prices.

The purpose of this paper is to analyze some cases, such as anticipated once-and-

---

*Department of Economics, Texas A&M University, College Station, TX 77843-4228. We are thankful to the members of the Macroeconomics Workshop at Texas A&M University and the two anonymous referees who provided very insightful help. The first author benefited from in any conversions with his graduate students and with G. Calvo on this topic.

[1] This condition was made clear in Guillermo A. Calvo (1977).

for-all changes in a state variable, in which the nonexistence of *unexploited* capital gains or losses at the time of the change does not necessarily imply continuity of the market or shadow price of assets, and in which such continuity cannot provide a general rule to be applied. There are important instances of economic problems falling into this category which have not been sufficiently explored in the literature, and for which we derive results not previously available. Moreover, as will be noted below, the method we propose is sufficiently general to handle the more conventional problems (such as the anticipation of changes in flow taxes or subsidies) in which the usual continuity of market or shadow prices is applicable.

In this paper we present the problem and suggest an intuitive, economic interpretation as a general method for dealing with it. The analysis generates results within a framework of familiar elementary price theory, and is done in the same spirit pioneered in Robert Dorfman's (1969) excellent economic interpretation of optimal control theory. The essence of the solution is to consider the optimal individual behavior "an instant" before the preannounced event is to take place, a procedure that is both extremely simple and sufficient to yield the whole path that we seek to investigate. In a recent paper (Leonardo Auernheimer, 1987) the same intuitive solution was derived heuristically and applied to a particular problem (an expected devaluation); here we both formalize the economic intuition and generalize the solution to a variety of other situations.[2]

In order to center the analysis on a familiar example, we present it in the context of the well-known one-sector neoclassical optimal growth model, in which we take an exogenous change of part of the physical capital stock as the future preannounced event.

I. The Problem

Take the case of the standard neoclassical one-sector growth model, with individuals who live forever and maximize the present value of utility given their constant rate of time preference. They hold a stock of the single kind of good which, combined with labor (for simplicity assumed here to be constant and normalized to unity), generates output which can be either consumed or accumulated. In other words, they seek to maximize

$$(1)\qquad \int_0^{\infty} e^{-\rho t}U(c_t)\,dt,$$

where c is consumption, ρ is the constant rate of time preference, and $U(c)$ is the instantaneous utility function, assumed to be strictly concave and twice differentiable, with first and second derivatives $U_c > 0$ and $U_{cc} < 0$. For simplicity we assume the capital stock depreciation rate to be zero. Therefore, we have

$$(2)\qquad dk/dt = f(k_t) - c_t,$$

where k is the per capita capital stock and $f(k)$ is the production function, which is assumed to be increasing, strictly concave, twice differentiable, and to satisfy $f(0) = 0$. The initial value of k is given by past history.

Under the assumption that the control variable, c, is piecewise continuous and right-hand continuous, we arrive at the well-known solution for the system, summarized by the set of two differential equations

$$(3)\qquad dc/dt = [U_c/U_{cc}][\rho - f'(k)]$$

$$(4)\qquad dk/dt = f(k) - c.$$

Consider, in this context, the effects of an announcement made at time $t = 0$ that at some future time $t = \tau > 0$ a part of the capital stock will be exogenously changed. Also, let the amount of capital which would be accumulated by τ in the absence of this announcement be denoted by $\hat{k}_\tau$, and let the amount of capital at the time of the announcement be denoted by k_0. (If at $t = 0$, k is the long-run steady state, then $\hat{k}_\tau = k_0$.) Before discussing the method of solution, the reader should notice that specific manner in which the capital stock will be changed is crucial for the result; after showing the general solution, we will consider three alternative specifications, all of which are relevant for particular situations and yield different insights.

---

[2] No rigorous mathematical proofs have been published to treat the problems covered intuitively in this paper. However, the Appendix to this paper sketches how to apply the rigorous proofs in the unpublished paper of Gabriel A. Lozada (1988), which is available from the author upon request.

II. The Nature of the Solution

The system (3)–(4) yields an infinite number of paths; for the standard case in which no changes are expected to occur in the future, the "saddle path" is the only one of those paths satisfying the transversality condition. In the case of our "experiment," though, the announcement opens up the possibility for (and the necessity of) one of the *other* feasible Euler paths to be the optimal one during the "interim" period between the announcement and the time the event takes place.[3] (After the event takes place, of course, no further changes will be forthcoming and the control variable—consumption, in this case—will adjust to its corresponding saddle path level.) To find which of those interim branches between $t = 0$ and $t = \tau$ should be selected, all we need to know are the values of consumption and capital just before and just after the change at τ. The reason for this is simple. Denote these values by c_τ^-, k_τ^-, c_τ^+, and k_τ^+. Then knowing c_τ^- and k_τ^-, one can run the differential equation system (3)–(4) backward in time to obtain the path for the interim period between $t = 0$ and $t = \tau$. Similarly, knowing c_τ^+ and k_τ^+, one can run the differential equation system forward in time to get the path for $t > \tau$. Therefore, the entire problem hinges on the four values of c_τ^-, k_τ^-, c_τ^+, and k_τ^+, and if a solution exists we need four equations to determine it. Three of these are obvious, but the fourth embodies the central insight of this paper.

For the first equation, we claim that starting from an initial capital stock k_0, the terminal value of consumption just before the change takes place is a continuous,[4] monotonically decreasing[5] function of the terminal value of the capital stock,

$$c_\tau^- = F_1(k_\tau^-). \tag{5}$$

For the second equation, notice that the saddle path toward the final steady state, after the preannounced event takes place, is uniquely defined by k_τ^+, the capital stock existing immediately after the event. In particular, the initial value of consumption on the saddle path after the event is a function of the initial stock of capital:

---

[3] It turns out to be easy but not trivial to formally prove that for $t \neq \tau$ paths in our problem obey (3)–(4), which come from problems with no shocks to k. These properties are rather obvious consequences of the method of proof in Lozada (1988).

[4] For the continuity of F_1, see (A30), p. 419 of Atle Seierstad and Knut Sydsæter (1987), on dependence on initial conditions of the solution of a differential equation. That result is applicable because the r.h.s. of (4) is twice differentiable.

[5] That the r.h.s. of (5) is a monotonically decreasing function means that given k_0 at $t = 0$, along any feasible Euler path a higher capital stock at a later date will require a lower value of consumption at that same date. This is not trivial, but the proof is straightforward. A sufficient condition for this to be the case is for any two continuous paths of consumption obeying (3)–(4) and starting with the same capital stock never to "cross" for any $0 < t < \tau^-$. To show that they never will, consider what would happen if two such paths $c_1(t)$ and $c_2(t)$, for which $c_1(0) > c_2(0)$, *did* cross for the first time at some t_1 between 0 and τ^-, so that $c_1(t_1) = c_2(t_1)$. Since c_2 would be crossing c_1 "from below," it would be true that $(dc_2/dt) > (dc_1/dt)$ at $t = t_1$, and from (3) it must be true that $k_1(t_1) \geq k_2(t_1)$, that is, that the capital stock corresponding to the path c_1 is *higher* than (or equal to) the capital stock corresponding to the path c_2 at $t = t_1$. But, on the other hand, since both consumption paths started with the same capital stock at $t = 0$, and since at all times from $t = 0$ to $t = t_1$, c_1 was higher than c_2, it follows that at $t = t_1$ the capital stock corresponding to the path c_1 must be strictly *lower* than the capital stock corresponding to the path c_2, which contradicts the previous statement. This shows that there can be no first crossing, which of course eliminates not only the possibility of single but also of multiple crossings. Notice that this proof is different than the proposition that in the phase diagram in (k, c) space no two feasible Euler paths ever cross. What the proof shows is that, starting from any initial capital stock, the attainment of a higher (lower) capital stock at some future time implies a lower (higher) level of consumption at all points during the adjustment.

$$c_\tau^+ = F_2(k_\tau^+). \tag{6}$$

It is trivial to show that F_2 is a monotonically increasing function, and continuity of F_2 follows directly from the fact that it describes the saddle path.

The third equation describes the specific form and magnitude of the exogenous change in the capital stock which occurs at τ,

$$k_\tau^+ = F_3(k_\tau^-). \tag{7}$$

In other words, F_3 describes a rule by which we can determine k_τ^+ simply by knowing k_τ^-. In all the cases we consider in this paper, we take F_3 to be nondecreasing. In other words, the more capital the individual has just before the shock, the more (or at least not less) capital he will own just after the shock.[6]

Until now this presentation has been completely rigorous. However, one more equation is needed to complete the system of four equations in four unknowns. The derivation of this equation is heuristic, and it proceeds as follows. We will first find the extent to which consumers are *willing* to trade c_τ^+ for c_τ^- at the margin. Then we will find the extent to which they are *able* to trade c_τ^+ for c_τ^- at the margin. Equating these two margins will yield the final equation.

Consider, in isolation, the situation of the typical individual facing the imminence of the "shock," and the evaluation, at $t = \tau - \Delta t$, of the sum of the utilities enjoyed both during a very small length of time immediately before and immediately after the date of the event. If those periods are sufficiently small, this sum of utilities can be written as

$$S = U(c_\tau^-)(\Delta t) + \frac{U(c_\tau^+)(\Delta t)}{(1+\rho)^{\Delta t}}, \tag{8}$$

where c_τ^-, c_τ^+ are the (constant) flows of consumption during those small periods, (Δt) is the length of each period, and, as before, ρ is the rate of time preference. Notice that although c_τ^- and c_τ^+ are strictly the flow levels "just before" and "just after" the change at $t = \tau$, in (8) and subsequent equations of this section they are meant to measure a constant flow of consumption during each of the two periods. Expression (8), then, is the utility enjoyed between times $(\tau - \Delta t)$ and $(\tau + \Delta t)$. Taking the total derivative of (8), equating the result to zero, and letting the length of the period approach zero yields the marginal necessary condition for a maximum,

$$\frac{U_c(c_\tau^-)}{U_c(c_\tau^+)} = -\frac{dc_\tau^+}{dc_\tau^-}. \tag{9}$$

Here dc_τ^+/dc_τ^- is the rate at which the consumer wishes to transform consumption in any one of the two small periods into consumption in the other period. Of course, this marginal rate of substitution has to be equated to the relevant marginal rate of transformation or "relative price" in the vicinity of τ. Thus we now need to investigate the consumer's capability to transform consumption from one period to the other (the budget constraint at the margin).

The key to specifying the individual's budget constraint around $t = \tau$ is to require that the capital stocks at $\tau - \Delta t$ and at $\tau + \Delta t$ remain fixed—that is, to require that whatever extra capital accumulation or disaccumulation takes place at the first period be completely compensated by disaccumulation or accumulation at the second period *after taking into account any change brought about by the exogenous shock.* This procedure allows us to disregard whatever happens outside the small neighborhood around $t = \tau$. In other words, we are asking the extent to which an extra unit of capital (and hence an extra unit consumption) can be "redistributed" or "transformed" between the two small periods within the neighborhood.

---

[6] We also assume that F_3 is continuous—in other words, that a small variation in the capital stock just before the shock will yield only a small variation in capital just after the shock. This is a natural assumption; for example, tax rules (which are one of the obvious policy analogues of "earthquakes") are in fact continuous. Finally, we assume that F_3 is differentiable, even though in one of the three particular cases considered in Section III below, F_3 has a kink point (where differentiability fails). As the reader will see, this causes no difficulties in the analysis; for a rigorous treatment, see the Appendix.

The constraint on capital accumulation is expressed as follows. We keep $k_{(\tau-\Delta t)}$ and $k_{(\tau+\Delta t)}$ fixed constants. Also, we treat c_τ^- and c_τ^+ as constant flows of consumption, and k_τ^- and k_τ^+ as constant levels of the capital stock, between $\tau-\Delta t$ and τ, or between τ and $\tau+\Delta t$, respectively. Since the effects of changes in the capital stock on production during either of the two periods are of the second order of smalls and can be ignored,[7] the constraints are

$$(10)\qquad k_\tau^- = k_{(\tau-\Delta t)} + [f(k_\tau^-) - c_\tau^-]\Delta t,$$

$$(11)\qquad k_\tau^+ = F_3(k_\tau^-),$$

$$(12)\qquad k_{(\tau+\Delta t)} = k_\tau^+ + [f(k_\tau^+) - c_\tau^+]\Delta t.$$

In order to find the rate at which consumption can be transformed between the two small periods we proceed to take the total derivative of expressions (10)–(12), keeping the length of the period constant. After a few simple substitutions we obtain

$$(13)\qquad \frac{(dc_\tau^+)\Delta t}{(dc_\tau^-)\Delta t} = -F_3'(k_\tau^-) \cdot \frac{1+f'(k_\tau^+)\Delta t}{1-f'(k_\tau^-)\Delta t}.$$

Letting the length of the period approach zero, the result is

$$(14)\qquad \frac{dc_\tau^+}{dc_\tau^-} = -F_3'(k_\tau^-).$$

Therefore, the rate at which marginal units of consumption can be transformed between the two periods is equal to the rate at which, at the margin, units of capital will be exogenously transformed at the time $t=\tau$. The simplest, trivial case is of course when there is no exogenous change in the capital stock at $t=\tau$, so that $F_3'=1$, and one unit of consumption in either of the two periods exchanges for one unit of consumption in other period.

Equating (9) and (14), we then obtain a simple but fundamental relationship,

$$(15)\qquad \frac{U_c(c_\tau^-)}{U_c(c_\tau^+)} = -\frac{dc_\tau^+}{dc_\tau^-} = F_3'(k_\tau^-).$$

If for convenience we define $l(k_\tau^-) = F_3'(k_\tau^-)$, then (15) can be written as

$$(16)\qquad \frac{U_c(c_\tau^-)}{U_c(c_\tau^+)} = l(k_\tau^-).$$

Of course, the same result can be obtained by directly maximizing (8) subject to the constraint (10)–(12).[8] Notice that the "shadow price" of a unit of capital is $\lambda = U_c(c)e^{-\rho t}$. Hence $\lambda_\tau^- = U_c(c_\tau^-)e^{-\rho\tau}$, $\lambda_\tau^+ = U_c(c_\tau^+)e^{-\rho\tau}$, and (16) becomes

$$(17)\qquad \frac{\lambda_\tau^-}{\lambda_\tau^+} = \frac{U_c(c_\tau^-)}{U_c(c_\tau^+)} = l(k_\tau^-).$$

The principle of "shadow price continuity" is therefore relevant only when $l(k_\tau^-)=1$. Of course, this will be the case if no shock is forthcoming.

We have now fully specified each equation in our system, but it remains to reduce the system of four equations into something more tractable. The first three equations (5)–(7) can easily be combined as follows. F_1 asserts that a higher k_τ^- implies a strictly lower c_τ^-, which implies a strictly higher $U_c(c_\tau^-)$. F_3 asserts that a higher k_τ^- implies a higher k_τ^+, which by F_2 implies a higher c_τ^+, and therefore a lower $U_c(c_\tau^+)$. Hence F_1, F_2,

---

[7] Using any k between $k_{(\tau-\Delta t)}$ and k_τ^- as the argument of f in (10), and using any k between k_τ^+ and $k_{(\tau+\Delta t)}$ as the argument of f in (12), changes nothing in (14), which is obtained after letting Δt go to zero.

[8] (16), and (16′) below, are the results which our heuristic technique was meant to produce. For a guide to a rigorous derivation, see the Appendix.

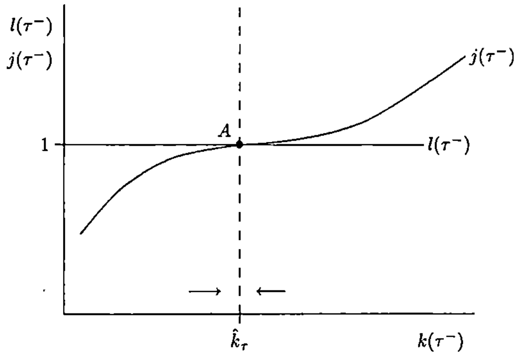


FIGURE 1

and F_3 imply that the ratio $U_c(c_\tau^-)/U_c(c_\tau^+)$ is a strictly increasing function of k_τ^-. We call this function j:

$$\frac{U_c(c_\tau^-)}{U_c(c_\tau^+)}=j(k_\tau^-). \tag{18}$$

Since F_1, F_2 and F_3 are all continuous, it is easy to see that j is continuous.

Substituting (16) into (18), we succeed in collapsing the system of four equations ((5), (6), (7), and (16)) in four unknowns (c_τ^-, k_τ^-, c_τ^+, and k_τ^+) into a system of one equation in one unknown,

$$l(k_\tau^-)=j(k_\tau^-). \tag{19}$$

Hence if we graph l and j, the point where they cross will determine the optimal k_τ^-.

In the graph of Figure 1, measure k_τ^- on the horizontal axis, and take $\hat{k}_\tau$ to be the capital stock which would have been accumulated by τ in the absence of any announcement.

On the vertical axis, measure two things. The first is the relative price $l(k_\tau^-)$, the l.h.s. of (19), which depends on the manner in which an eventual preannounced change is to take place—that is, on the rule specifying how every unit of capital will be changed as it "crosses over" the time point $t=\tau$ (a situation quite similar to the imposition of a tariff or the granting of a subsidy on a commodity as it "crosses over" a territorial border). The second thing we measure on the vertical axis is the ratio of the marginal utilities just before and just after the change, $j(k_\tau^-)$, which is the r.h.s. of (19); $j(k_\tau^-)$ is upward sloping.

Now, notice what happens if at $t=0$, with an initial capital stock k_0 and a well-defined announcement of whatever change will take place at $t=\tau$, any given "target" k_τ^- is being considered instead of $\hat{k}_\tau$ as a possible optimum. If at the k_τ^- being considered the r.h.s. (j) of (19) is larger than the l.h.s. (l) of (19), then it is easy to show that a small decrease of the target k_τ^- will increase welfare, and vice versa. This is indicated in Figure 1 and following by horizontal arrows.

Figure 1 describes the simplest case in which *no change* will take place at $t=\tau$, so that $l=1$ and $k_\tau^-=k_\tau^+$. Here, of course, the optimal strategy is for consumers to remain at $\hat{k}_\tau$, and therefore $j(k_\tau^-)=l(k_\tau^-)=1$. Although for this trivial case the representation in Figure 1 does not add much to what can be found by simple inspection of (16), the same little apparatus will prove to be useful in the interpretation of various possible preannounced changes.

To conclude this section, consider how the analysis would proceed if in addition to the preannounced change in the state variable, k, there had been a preannounced imposition of a tax on the flow of consumption t_c and a tax or subsidy on net investment t_i. In this case, the same reasoning from (10) to (16) would have yielded an expression equal to[9]

$$\left[\frac{1+t_c}{1+t_i}\right]\frac{U_c(c_\tau^-)}{U_c(c_\tau^+)}=l(k_\tau^-). \tag{16$'$}$$

The "shadow prices" of a unit of capital would be $\lambda_\tau^-=U_c(c_\tau^-)e^{-\rho\tau}$ and $\lambda_\tau^+=((1+t_i)/(1+t_c))U_c(c_\tau^+)e^{-\rho\tau}$, so from (16′) it is immediate that

$$\frac{\lambda_\tau^-}{\lambda_\tau^+}=\left[\frac{1+t_c}{1+t_i}\right]\frac{U_c(c_\tau^-)}{U_c(c_\tau^+)}=l(k_\tau^-). \tag{20}$$

---

[9] To check this, note that while (10) and (11) are unchanged, (12) becomes $k_{(\tau+\Delta t)}=k_\tau^+ +[f(k_\tau^+)-(1+t_c)c_\tau^+]\Delta t-t_i[k_{(\tau+\Delta t)}-k_\tau^+]$.

From (20) it is clear that if there is no shock to capital (and therefore $l(k_\tau^-)=1$), then the problem can be easily handled by the shadow price continuity rule. In such cases, even though λ_τ^- will equal λ_τ^+, there will obviously still be a discontinuity in c at τ (except when $t_i = t_c$). These are the more conventional cases referred to in the introduction and considered, for example, by Kenneth L. Judd (1985a, b, 1987).

III. Some Applications

As indicated before, we now proceed to apply the solution to three possible cases:

(i) The capital stock changes at $t=\tau$ by a given percentage $z=(\Delta k/k)$.

(ii) At $t=\tau$ the capital stock *of each individual*[10] is changed by a fixed absolute amount (Δk).

(iii) At $t=\tau$ either *each individual's* capital stock is "leveled down" to a given level k' (i.e., all units in excess of k' are to be annihilated) or "leveled up" to a given level k' (i.e., the individual's stock will be raised, if below a given level k', to that level). Notice that these are two different subcases; later on we will comment on the possibility of both these two changes being imposed simultaneously.

Each of these cases has a correspondence to a relevant economic problem. The first one is similar, when $z<0$, to an expected devaluation of z percent in a model in which, in the absence of capital mobility, the "capital stock" is the stock of real money balances.[11] The second is the anticipation of a once-and-for-all absolute tax (or subsidy) on the ownership of capital. The third case (which could in general be characterized as an "equalizing policy") has two subcases. They correspond, respectively, to: (a) the anticipation of a "social" reform in which private ownership of capital above a certain maximum is confiscated (though our example allows for free accumulation subsequently to the change at $t=\tau$); and (b) to the anticipation of "free capital distribution." Perhaps the first could be called a "downward equalizing" and the second an "upward equalizing" policy. We consider each one of these cases in turn.

(i) Consider, first, the case in which at $t=\tau$ the capital stock will exogenously change by a fixed percentage z. The "rate of transformation" or "exchange" between dc_τ^- and dc_τ^+, is, in this case,

$$dc_\tau^+ = -\,dc_\tau^-[1+z] \tag{21}$$

from (14), so that (16) can be written as

$$\frac{U_c(c_\tau^-)}{U_c(c_\tau^+)} = 1+z. \tag{22}$$

If $z<0$ than (22) implies $U_c(c_\tau^+) > U_c(c_\tau^-)$; concavity of U then implies that $c_\tau^+ < c_\tau^-$: the level of consumption will fall discretely at the time of the shock. For a more explicit statement of this, recall that by the Mean Value Theorem

$$U_c(c_\tau^+) - U_c(c_\tau^-) = U_{cc}(c)(\Delta c),$$

where $(\Delta c) = c_\tau^+ - c_\tau^-$ is the change ("jump") of the level of consumption at $t=\tau$, and where c is between c_τ^+ and c_τ^-. Then the requirement (22) can be written as

$$\frac{U_{cc}(c)(\Delta c)}{U_c(c_\tau^+)} = -\,z, \tag{22'}$$

so that the "jump" in consumption to take place at $t=\tau$ is given by

$$\Delta c = \frac{-\,zU_c(c_\tau^+)}{U_{cc}(c)} < 0 \text{ for } z<0. \tag{23}$$

Notice that (17) and (22) imply for this case that the ratio of the "shadow price" of capital before and after the shock is given by

---

[10] We wish to avoid the "free-rider" problem that would arise in cases (ii) and (iii) if an absolute part of the *aggregate* capital stock (rather than a part of every *individual's* capital stock) were to be destroyed.

[11] This particular case is analyzed in Auernheimer (1987). See also Calvo (1986) and Maurice Obstfeld (1985).

$$\frac{\lambda_\tau^-}{\lambda_\tau^+} = 1 + z. \tag{24}$$

Clearly, then, the shadow price continuity principle does not hold. However, since λ is the shadow price of a unit of capital, we can interpret $V = \lambda k$ as the "shadow value" of the capital stock. Then since $k_\tau^+/k_\tau^- = 1 + z$ by assumption, from (24) we have

$$\frac{V_\tau^-}{V_\tau^+} = \frac{\lambda_\tau^- \cdot k_\tau^-}{\lambda_\tau^+ \cdot k_\tau^+} = \frac{1+z}{1} \cdot \frac{1}{1+z} = 1.$$

Therefore, the solution dictates that in this case no discontinuity is to occur in the shadow value λk at the time of the foretold exogenous shock, even though there is a discontinuity in the shadow price λ. A slightly different interpretation, which generalizes the notion of shadow price continuity, is that from (24) we can write $\lambda_\tau^- = (1+z)\lambda_\tau^+$. That is, continuity of the "net" shadow price is preserved, in very much the same manner in which in standard price theory "before tax" and "after tax" prices are related to each other.

Figure 2 shows the application of the functions $l(\cdot)$ and $j(\cdot)$. In this case, the change preannounced at $t = 0$ determines a fall in l, the relative price of capital before and after τ, from 1 to $(1+z)$. The function $j(k_\tau^-)$ shifts downward: the relationship between k_τ^- and c_τ^- (and hence the numerator in the l.h.s. of (18)) is unchanged, but now k_τ^+ (and hence c_τ^+) is less than in Figure 1 (for which $k_\tau^+ = k_\tau^-$), so for any given k_τ^- the denominator in the l.h.s. of (18) is larger. Although Figure 2 describes an equilibrium at point A with $k_\tau^- < \hat{k}_\tau$ – that is, less capital was accumulated than if there had been no announcement—whether more or less accumulation will occur depends on the relative magnitude of those shifts, and nothing can be said *a priori*.[12]

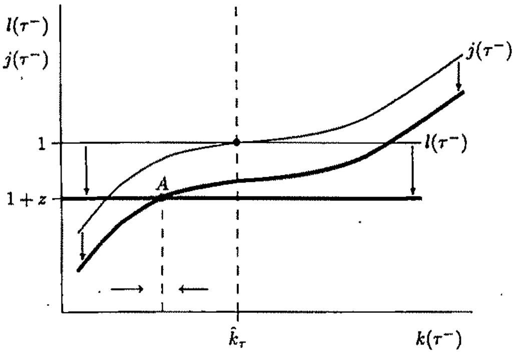


FIGURE 2

(ii) The second case to consider is when the "shock" consists of a fixed absolute change in the capital stock of every individual. In this case, no matter what the level of the capital stock is at the time of the shock, a unit of capital from any of the two short periods before and after the shock will be "exchanged" for, or "transformed" into, exactly one unit in the other period. In this case (15) becomes

$$\frac{U_c(c_\tau^-)}{U_c(c_\tau^+)} = -\frac{dc_\tau^+}{dc_\tau^-} = 1$$

and (17) takes the form

$$\lambda_\tau^+/\lambda_\tau^- = 1, \tag{25}$$

which is the same as the standard result of no discontinuities occurring in the shadow price of capital, and hence in the level of consumption. The intuitive interpretation of the result is immediate: since the absolute level of the shock is unrelated to the existing stock of capital at the time of shock, the "exchange ratio" (terms of trade) of marginal units of capital between one instant and the next is the same as in the absence of the shock.

The graphical interpretation of this case is provided in Figure 3. There is no change in the relative price, $l(\cdot)$, which remains at unity, but the curve $j(\cdot)$, the ratio of marginal utilities, falls, for the same reason as in Figure 2, namely that, for any given k_τ^-, k_τ^+ is less than it was before the an-

---

[12] An analysis of this point is given in Auernheimer and Lozada (1989).

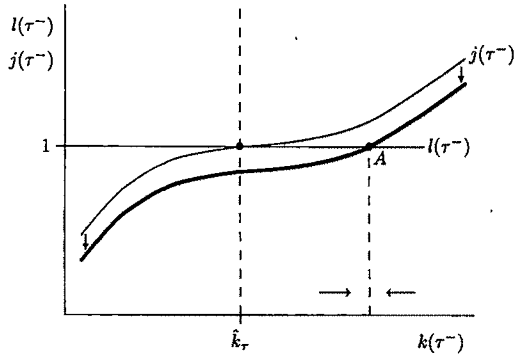


FIGURE 3

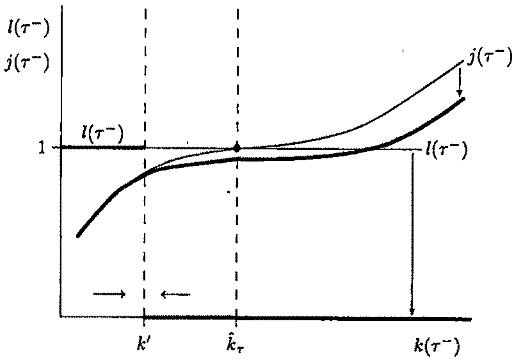


FIGURE 4

nouncement. The new intersection of $l(\cdot)$ and $j(\cdot)$ now determines an optimal point for k_τ^- above the previously planned level, $\hat{k}_\tau$. In other words, more capital accumulation takes place during the interim period from $t=0$ to $t=\tau$ than if no announcement had been made.

(iii) The third case, as suggested before, encompasses two alternative "subcases." In both these subcases, our graphical analysis will determine what happens to k at τ; what happens to consumption c at τ will be shown in the Appendix.

The first subcase is the one for which the anticipated shock is to be the "annihilation" of all of the individual's capital stock in excess of a fixed level k'. Of course, if $k' \geq \hat{k}_\tau$ then the optimal response is to continue to plan on reaching $\hat{k}_\tau$, and nothing happens, so we need to take $k' < \hat{k}_\tau$ to make the problem interesting. In this case the constraint (14) takes the form

$$\text{(26a)} \quad -dc_\tau^+/dc_\tau^- = 0 \begin{cases} \text{for } k_\tau^- > k', \text{ or} \\ \text{for } k_\tau^- = k' \text{ and } dk_\tau^- > 0 \end{cases}$$

$$\text{(26b)} \quad -dc_\tau^+/dc_\tau^- = 1 \begin{cases} \text{for } k_\tau^- = k' \text{ and } dk_\tau^- < 0, \text{ or} \\ \text{for } k_\tau^- < k'. \end{cases}$$

What (26a) and (26b) dictate is a corner solution. For $k_\tau^- > k'$, the cost of increasing consumption at τ^- is zero, so that (16) cannot hold for a finite c_τ^-. The solution then needs to be one for which the capital stock just before the change, k_τ^-, is exactly equal to k': at such a point it does not pay to accumulate, because all accumulation will be eliminated at the next instant, or disaccumulate, since the right-hand side of (16) becomes unity for $dk_\tau^- < 0$. If $k_\tau^- = k'$, of course, no "annihilation" actually occurs.[13]

For this first subcase, Figure 4 provides the graphical counterpart of the results. Notice that both the curve $j(\cdot)$, representing the ratio of marginal utilities, and the relative price, $l(\cdot)$, change only to the right of the given critical value k'. If $k_\tau^- > k'$, then an additional unit of capital given up just before the change at $t=\tau$ will exchange for nothing, so that $l=0$. Furthermore, for any given $k_\tau^- > k'$ the level of capital after the change will be equal to k'—that is, it will be smaller than in the initial position in which no change is expected, so that j falls. This determines a new optimal k_τ^- at k', which implies less capital accumulation than before. Although for $k_\tau^- = k'$, the optimal terminal level of the capital stock chosen to be reached at τ^-, there is no well-defined relative price l, this price can be thought of as being given by the ratio of marginal utilities at that point, that is, by the height of curve $j(\cdot)$. This is the exact analogue of the long-run static equilibrium position of consumers faced with a "kinked" production possibilities frontier and an equilibrium at the kink point, where the relative price is given exclusively by demand conditions. Indeed, as is shown in the Appendix (compare Proposition A1), $U_c(c_\tau^-)/U_c(c_\tau^+)$ is less than one here, which implies that consumption falls at the time of the shock, and that the shadow price continuity principle does not hold.

---

[13] The outcome resembles the case of the train commuter who, intrigued by a fellow passenger's behavior, inquires "Why do you tear your newspaper apart every evening and throw the little pieces out of the window as we enter New Jersey?"

—To scare the tigers.

—But there are no tigers in New Jersey...

—Now you know why.

The second subcase is the mirror image of the first: the announcement is that at $t = \tau$ every individual's capital stock, if smaller than some arbitrary value k', will be increased to k'. The relevant conditions are, then,

$$\text{(27a)}\quad -dc_\tau^+/dc_\tau^- = 0\begin{cases}\text{for } k_\tau^- < k', \text{ or}\\ \text{for } k_\tau^- = k' \text{ and } dk_\tau^- < 0\end{cases}$$

$$\text{(27b)}\quad -dc_\tau^+/dc_\tau^- = 1\begin{cases}\text{for } k_\tau^- = k' \text{ and } dk_\tau^- > 0, \text{ or}\\ \text{for } k_\tau^- > k'.\end{cases}$$

The first, trivial possibility is for the arbitrary level k' to be the same or higher than the planned level $\hat{k}_\tau$; in this case (27a) becomes relevant, and the capital stock is completely consumed by the time of the change, so that $k_\tau^- = 0$. In the more interesting case in which $k' < \hat{k}_\tau$, either the capital stock is totally exhausted, so that $k_\tau^- = 0$, or nothing happens, so that $k_\tau^- = \hat{k}_\tau$. Proposition A3 in the Appendix shows that for a given planned capital stock $\hat{k}_\tau$ there is a "trigger value" of k', between zero and $\hat{k}_\tau$, below which the announcement will have no effect and no "gift" will therefore take place, and above which the total exhaustion of capital will occur.

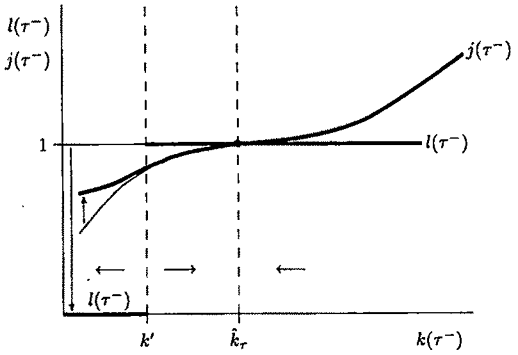


FIGURE 5

For this second subcase the graphical interpretation is given by Figure 5. In this subcase, the counterpart of the first one, changes in the curve $j(\cdot)$ and the cost line $l(\cdot)$ occur only to the left of $k_\tau^- = k'$. If $k_\tau^- < k'$, then an additional unit not consumed just before the change will result in no gain in units just after the change, so that the "exchange ratio" l becomes zero. Also, for $k_\tau^- < k'$, to any k_τ^- there will correspond a higher k_τ^+ than before, and therefore a higher value of the $j(\cdot)$ function. Here, two outcomes are possible: either nothing happens, if k' is close enough to zero, or the capital stock is completely depleted, with $k_\tau^- = 0$, if k' is close enough to $\hat{k}_\tau$. This is indicated by the horizontal arrows; if any $0 < k_\tau^- \leq k'$ is being considered, then a better option is a smaller value of k_τ^- (as any additional unit consumed before $t = \tau$ will be replaced at no cost); at the same time, for any $\hat{k}_\tau > k_\tau^- > k'$, a higher k_τ^- is a better candidate (the minimum k_τ^- to qualify for the "gift" is not attained, and there is no more incentive to depart from $\hat{k}_\tau$ than in the absence of any announcement). As shown in the Appendix, then, for a given τ there is a "trigger value" of k' that will determine which of the two extreme outcomes occurs.

As in the first subcase of case (iii), the behavior of consumption and the shadow price at τ is a topic for the Appendix (Proposition A2). There it is shown that, if it turns out to be optimal to receive the gift, shadow

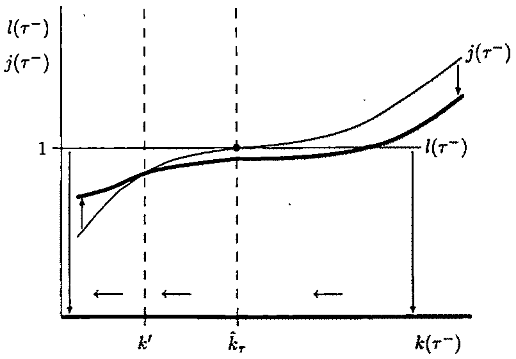


FIGURE 6

price continuity would hold only by coincidence.

For the sake of completeness it is interesting to notice that a third possibility is the simultaneous occurrence of the two previous subcases, that is, the case in which at $t=\tau$ the individual's capital stock will be decreased to k', if higher than k', and increased to k', if lower than it —perhaps a "two-way equalizing" policy could be an adequate name. In this case

$$dc_\tau^+/dc_\tau^- = 0$$

for all k_τ^-, and complete exhaustion of the capital stock at $t=\tau$ always occurs. The case combining the first and the second subcases is shown in Figure 6, which needs no elaboration.

IV. Concluding Remarks

We have presented a heuristic method, based on elementary notions of price theory, for analyzing the result of anticipated changes in models of dynamic optimization with perfect foresight. The method is very general and can be applied to any anticipated change. However, we have discussed it in the context of exogenous changes in the state variable because in this case the "shadow price continuity principle" can fail —and hence it is in this case that the method provides results not previously available. We also applied the method to a variety of cases, each corresponding to instances of possible anticipated policy changes, illustrating each case with a simple graphical device.

Before ending, three final remarks are in order.

First, notice that our results depend only on "local" conditions and are independent, for example, of the rate of time preference or the particular form of the production function.

The second remark concerns existing work which considers the case of anticipated parameter changes, such as the imposition of flow taxes. As we showed at the end of Section II, our discussion can accommodate such changes, but these are cases in which the shadow price continuity principle applies. There is a discontinuity in the level of consumption at the time the preannounced event takes place, but the discontinuity is a direct consequence of the change in the parameters at that time. In fact, it is precisely the continuity of the shadow price that necessitates the discontinuity in the level of consumption.

The third remark is that when there are anticipated changes in the state variable, the *only* case in which shadow price continuity must obtain is when the magnitude of the change is *totally unrelated* to the stock of capital before the change. At first glance, this is a rather surprising result. Recall, though, that the shadow price is an endogenous variable, whose level serves to influence the optimal consumption path. If marginal changes in consumption prior to the shock can influence the magnitude of the shock, then one would expect the shadow price to signal this. Immediately after the shock the shadow price no longer needs to signal the imminence of a shock, and therefore jumps. If, on the other hand, marginal changes in consumption prior to the shock cannot affect the magnitude of the shock, then there is no reason for the shadow price to signal differently just before and just after the shock. Of course, such "lump sum" preannounced changes will in general have an effect on the *level* of the shadow price at the time of the shock, but there will be no discontinuity at that time.

APPENDIX

Of the two basic curves $j(k_\tau^-)$ and $l(k_\tau^-)$ which were used in the body of this paper to provide a complete solution to the problem, $j(k_\tau^-)$ was derived in a fully rigorous way. (It did not use the differentiability of F_3). Therefore, (18) holds. The first task of this Appendix, then, is to sketch how to rigorously verify the results which in the text are based on $l(k_\tau^-)$—that is, results which come from the discussion beginning with (8) and culminating with (16).

For cases (i) and (ii), this verification is a consequence of the results derived in Lozada (1988), which is available upon request. That paper shows that if F_3 is differentiable, then under a mild technical assumption,[14] $\lambda_\tau^-/\lambda_\tau^+ = F_3'$. This is true even if announced flow taxes on consumption and net investment are imposed at τ. To get (16), (16′), and (20) from this, all one has to do is maximize the relevant Hamiltonian.

Figure 4 suggests that in the first subcase of case (iii), if $k' < \hat{k}_\tau$ then the optimal k_τ^- is k'. (If $k' \geq \hat{k}_\tau$ then one has the trivial case in which neither the announcement nor the shock have any effect). Because F_3 is not differentiable at k' in this subcase, the result mentioned in the last paragraph, $\lambda_\tau^-/\lambda_\tau^+ = F_3'(k)$, holds only for $k \neq k'$. The characteristics of the optimal path are derived as follows.

PROPOSITION A1: *For the first subcase of case (iii), if* $k' < \hat{k}_\tau$ *then* $k_\tau^- = k'$, $c_\tau^- > c_\tau^+$, *and* $\lambda_\tau^- < \lambda_\tau^+$.

PROOF:

It is obvious that $k_\tau^- \leq k'$. Then from (18), $U_c(c_\tau^-)/U_c(c_\tau^+) = j(k_\tau^-) \leq j(k') < j(\hat{k}_\tau) = 1$, which yields $c_\tau^- > c_\tau^+$. The basic relation $\lambda_t = U_c(c_t)\,e^{-\rho t}$ then implies $\lambda_\tau^- < \lambda_\tau^+$.

To prove that $k_\tau^- \leq k'$ is satisfied by equality, suppose instead that $k_\tau^- < k'$. Then F_3' is well-defined and equals one, and therefore $\lambda_\tau^- = \lambda_\tau^+$, which is false. □

The second subcase of case (iii) is the only case remaining, but it is the hardest one of the paper. First of all, one has the basic behavior of k and c:

PROPOSITION A2: *For the second subcase of case (iii), if it is optimal not to become eligible for the gift, then* $k_\tau^- = k_\tau^+ = \hat{k}_\tau$, $\lambda_\tau^- = \lambda_\tau^+$, *and* $c_\tau^- = c_\tau^+$. *On the other hand, if it is optimal to become eligible for the gift, then* $k_\tau^- = 0$, $\lambda_\tau^- \geq 0$, *and* $c_\tau^- > 0$, *while* $k_\tau^+ = k'$, $\lambda_\tau^+ > 0$, *and* $c_\tau^+ > 0$.

PROOF:

The first sentence is obvious because in this case nothing happens. For the second sentence, $k_\tau^- = 0$ is clearly optimal (the proof is left to the reader). Hence $F_3(k_\tau^-) \equiv k'$ and $F_3' = 0$. Taking account of the constraint $k_\tau^- \geq 0$, one has $\lambda_\tau^-/\lambda_\tau^+ \geq F_3'$, with equality if $k_\tau^- > 0$; therefore, only $\lambda_\tau^-/\lambda_\tau^+ \geq 0$ is necessary here. c_τ^- (and from it, λ_τ^-) is uniquely determined by (3)–(4) and the boundary conditions $k(0) = k_0$ and $k(\tau) = 0$ (compare Akira Takayama, 1985, p. 623). Finally, since (k_τ^+, c_τ^+) is on the saddle path, it is obvious that c_τ^+ is finite and positive and hence that $\lambda_\tau^+ > 0$. □

This justifies the claims made in the paper concerning the behavior of c at τ. The final task of this Appendix is to verify the claim that there is a "trigger value" (call it κ) of k'.[15]

PROPOSITION A3: *There is a unique number* κ, *lying between* 0 *and* $\hat{k}_\tau$, *such that the following is true. If* $k' < \kappa$ *then it is optimal not to become eligible for the gift. If* $k' > \kappa$ *then it is optimal to become eligible for the gift. (If* $k' = \kappa$ *one is indifferent between becoming eligible for the gift or not).*

PROOF:

Let $\hat{V}$ be the value of the "no-gift path" $\hat{k}_t$. Let V_g be the optimal value of the program which has $k_\tau^- = 0$. From Proposition A2, if the gift is given then V_g is the optimal value of the objective function.

Define ΔV to be $V_g(k') - \hat{V}$, where we write k' as an argument of V_g to emphasize the dependence of V_g on k'. We wish to show that there exists a unique $\kappa \in (0, \hat{k}_\tau)$ such that $\Delta V < 0$ for all $k' < \kappa$, $\Delta V = 0$ for $k' = \kappa$, and $\Delta V > 0$ for all $k' > \kappa$. One can use the Intermediate Value Theorem of elementary calculus to show this, provided that $\Delta V(k')$ is: (a) continuous; (b) monotonic; (c) has $\Delta V(0) < 0$; and (d) has $\Delta V(\hat{k}_\tau) > 0$. These four properties are proven as follows. (Uniqueness follows from monotonicity.)

Continuity: Break $V_g(k')$ up into two parts V_{g1} and $V(k')$, the first representing the contribution to the value function of c_t, $t \in [0, \tau]$, and the second representing the contribution to the value function of c_t, $t > \tau$. Then V_{g1} is not a function of k': it is simply the value of the program which optimally drives the capital stock to zero at τ. The problem corresponding to $V(k')$ is merely the standard problem of maximizing $\int_\tau^\infty e^{-\rho t} U(c_t)\,dt$ subject to the usual constraints (see Section I) and $k_\tau = k'$. Under the assumptions which have been made, $V(k')$ is differentiable, from (135) of Theorem 9, p. 213 of Seierstad and Sydsæter (1987). Thus it is continuous.

Monotonicity: Since $V_g(k') = V_{g1} + V(k')$, it is obvious that $V_g(k')$ is monotonically increasing, since $V(k')$ is monotonically increasing.

$\Delta V(0) < 0$: If $k' = 0$ then the size of the gift is zero (so no gift is ever made). So if $k' = 0$, the *problems* corresponding to V_g and to $\hat{V}$ are the same. Hence $V_g(0) < \hat{V}$, since $\hat{V}$ is optimal for this problem by definition. (We know that $V_g(0) \neq \hat{V}$ because: (i) the optimal k_t is unique,[16] and (ii) because the problems corresponding to V_g and $\hat{V}$ do not share the same optimal k_t since $\hat{k}_t$ is an interior program).

---

[14] Denote by (P_2) the problem of maximizing $V(k_\tau^{II})$, where $V(k_\tau^{II})$ is defined to be $\int_\tau^\infty e^{-\rho t} U(c_t)\,dt$, subject to all the conditions of Section I, and subject to k_τ^{II} exogeneously fixed; k_τ^{II} means the value of k_τ in problem (P_2). Then assume that an optimal solution exists for the problem (P_2) at $k_\tau^{II} = k_\tau^+$, and that $V(k_\tau^+)$ is well defined for k in some neighborhood of k_τ^+.

[15] Proposition A3 requires the following technical assumption. Define (P_2) as in fn. 14. Then assume that an optimal solution exists for (P_2) at k', and $V(k')$ is well defined for k in some neighborhood of k'.

$\Delta V(\hat{k}_\tau) > 0$: If $k' = \hat{k}_\tau$ then $V_g > \hat{V}$, since consumption for the two programs will be identical after τ, and consumption will be greater before τ in the program corresponding to V_g, which drives the capital stock to zero. ■

REFERENCES

Auernheimer, Leonardo, "On the Outcome of Inconsistent Programs under Exchange Rate and Monetary Rules," *Journal of Monetary Economics*, March 1987, *19*, 279–305.

______ **and Lozada, Gabriel A.** "To Announce or Not to Announce?," "An Analysis in the Context of the Theory of Public Finance," Unpublished manuscript, Department of Economics, Texas A&M University, 1989.

Calvo, Guillermo A., "The Stability of Models of Money and Perfect Foresight: a Comment," *Econometrica*, October 1977, *45*, 1737–39.

______, "Temporary Stabilization: Predetermined Exchange Rates," *Journal of Political Economy*, December 1986, *94*, 1319–29.

Dorfman, Robert, "An Economic Interpretation of Optimal Control Theory," *American Economic Review*, December 1969, *59*, 817–31.

Judd, Kenneth L. (1985a), "Redistributive Taxation in a Simple Perfect Foresight Model," *Journal of Public Economics*, October 1985, *28*, 59–83.

______, (1985b), "Short-Run Analysis of Fiscal Policy in a Simple Perfect Foresight Model," *Journal of Political Economy*, April 1985, *93*, 298–315.

______, "The Welfare Cost of Factor Taxation in a Perfect-Foresight Model," *Journal of Political Economy*, August 1987, *95*, 675–709.

Lozada, Gabriel A. "Optimal Control with Anticipated Shocks to the State Variables," unpublished manuscript, Department of Economics, Texas A&M University, 1988.

Obstfeld, Maurice, "The Capital Inflows Problem Revisited: A Stylized Model of Southern Cone Disinflation," *Review of Economic Studies*, October 1985, *52*, 605–25.

Sargent, Thomas and Wallace, Neil, "The Stability of Models of Money and Growth with Perfect Foresight," *Econometrica*, November 1973, *41*, 1043–48.

Seierstad, Atle, and Sydsæter, Knut, *Optimal Control Theory with Economic Applications*, Amsterdam: North-Holland, 1987.

Takayama, Akira, Mathematical Economics, 2nd ed., Cambridge: Cambridge University Press, 1985.

---

[16] Uniqueness comes from the Mangasarian sufficiency theorem (Seierstad and Sydsæter (1987, pp. 105–106) and the strict concavity of the Hamiltonian in (k, c).

Inventories and the Propagation of Sectoral Shocks

By RUSSELL COOPER AND JOHN HALTIWANGER*

This paper contrasts the dynamic properties of an imperfectly competitive economy with a representative agent, real business cycle model. For both economies, inventories are the important dynamic linkage. The predictions of these models with regards to the comovement of employment across sectors may differ. Empirical evidence on the comovement of employment over the business cycle is consistent with the model of imperfect competition with inventory holdings. (JEL 023,131)

Macroeconomists are generally concerned with characterizing the positive and normative aspects of business cycles. Two important features of cycles are: (i) correlated movements in aggregate output over time and (ii) the positive comovement in output and employment between sectors over the cycle.[1] For the purpose of building models of the cycle, the task is to determine the features of preferences, endowments, technology, and information, along with the institutional structure in which agents interact, that reproduce these two key observations.

Macroeconomic theorists have focused on correlated movements in output over time by studying the wide variety of intertemporal linkages (capital, non-separable preferences, stochastic structure) present in the economy. This emphasis on persistence reflects both the extensive use of the representative agent paradigm and the fundamental concern of macroeconomists over aggregate quantities. This literature highlights both possible sources of aggregate fluctuations (monetary shocks, real shocks, sunspots) as well as the mechanisms which propagate these disturbances throughout the economic system.

Formulating models to understand the comovement in employment and output across sectors is a bit more difficult. To see why, consider an extreme economy in which workers suffer no disutility from work so that equilibrium is always at full employment. Then, variations in output and employment in one sector, say due to sector specific technology shocks, will cause opposite movements in other sectors. That is, sectoral shocks may produce the *substitution* of resources from one sector to another rather than the positive comovement observed in practice. While these substitution effects dominate in this extreme case, they are certainly present in other, less-extreme, settings as well.[2]

There are a number of alternatives to this extreme model which will generate positive comovement in employment and output across sectors of an economy. Robert Lucas (1972,1975) accomplishes this in an economy with aggregate monetary disturbances which

---

*Department of Economics, University of Iowa and Department of Economics, University of Maryland, respectively. Helpful suggestions and comments from Steve Davis, Joel Fried, John Taylor, participants at the 1987 NBER Summer Institute Meetings, seminar participants at the Federal Reserve Bank of San Francisco, University of Maryland, the University of Rochester and two anonymous referees are greatly appreciated. Financial support from the National Science Foundation for Cooper, through SES-8605302 and a grant to the Hoover Institution's National Fellows Program, and for Haltiwanger, through SES-8520697, is gratefully acknowledged.

[1] Discussions of these observations appear, for example, in Robert Lucas (1977), John Long and Charles Plosser (1983, 1987) and Victor Zarnowitz (1985). We provide empirical evidence in Section IV of this paper on employment fluctuations.

[2] There is a literature on sectoral flows of workers in the presence of sector-specific shocks. See, David Lilien (1982) and Katharine Abraham and Lawrence Katz (1986) for a discussion of these flows and their relation to aggregate economic activity. Richard Rogerson (1987) discusses sectoral flows in a two-period model with sector specific shocks. His model implies negative correlations in employment levels as workers move from one sector to another.

individual agents confuse with real shocks. As an alternative, aggregate real disturbances could be the source of aggregate fluctuations in employment and output. An increase in economywide productivity would generally increase output in all sectors of the economy. John Long and Charles Plosser (1983) present a model in which aggregate fluctuations in output arise from sector specific rather than aggregate technology shocks. This emphasis on sectoral shocks is perhaps more convincing since it is difficult to identify aggregate real disturbances associated with each turn of the business cycle. In the Long-Plosser model economy, a technology shock in one sector leads to a contemporaneous expansion in that sector's output. This additional output is partly consumed and partly used as an input in the production of other commodities for *next* period.[3] The resulting output of the other commodities is then partly consumed and partly used as additional inputs, etc. In this manner shocks to the production function in a single sector are spread over time and to other sectors creating persistent aggregate fluctuations. Because of their assumptions on preferences, employment levels do not fluctuate in the Long-Plosser formulation. We return to this point later.

This paper provides an alternative framework for understanding the comovement in output *and* employment in a multisector setting. In contrast to Long-Plosser's emphasis on technological linkages and *factor* demand flows, our approach highlights demands for *final* consumption goods as the important linkage between sectors and excludes production of commodities by commodities. Instead, the intertemporal linkage in the model is the holding of inventories by the firm in one sector and the linkage across sectors is through demands for final goods only. Further, we are interested in generating empirical implications for the comovement in employment levels across sectors over the cycle.

The normality of demands for consumption goods is the basis for the propagation of shocks from one sector to another. These demand linkages create output and employment movements rather than price changes when the economy is in an underemployment equilibrium in which all available resources are not fully utilized. The source of the underemployment is imperfect competition.[4] Oliver Hart (1982), Martin Weitzman (1985), Cooper (1986) and Robert Hall (1988) also emphasize the importance of imperfect competition in generating large quantity fluctuations relative to perfect competition because the economy is more likely to be in an underemployment region when sellers of goods and/or labor have market power.[5]

One important feature of our model is that only *one* sector can hold inventories. This reflects the fact that many sectors of the economy produce goods and services which cannot be held in inventory. However, as a consequence of the demand linkages across sectors, the holdings of inventories in a subset of the sectors is sufficient to produce persistence in the output of all sectors. A buildup of inventories in one sector will reduce production in that sector which then leads to a reduction in demand for the product of other sectors. In the underemployment equilibrium this decrease in demand is met by reductions in output and employment. Section I presents a version of this model in which imperfect competition is a key feature of the environment and validates these claims about persistence and comovement.

To facilitate a comparison of approaches to studying the comovements and persistence of output and employment, Section II presents a representative agent, competitive real business cycle model with inventory holdings in one sector. Absent mobility costs, shocks to this system will produce negatively correlated movements in output and employment as labor time is shifted from one sector to another. As one sector expands, the other tends to contract.

---

[3] Thomas Sargent (1981) stresses similar factor demand linkages in his discussion of a corn-hog model.

[4] See Robert Hall (1986) for a lengthy discussion of the macroeconomic implications of imperfect competition.

[5] Strictly speaking, imperfect competition is not necessary for the effects discussed here to occur. As discussed in Section III, the market power of firms increases the probability that the economy will be in an underemployment region.

Section III reconciles these results with the previous model of imperfect competition. This discussion highlights the role of market structure and the heterogeneity of agents' preferences in determining whether employment movements are positively or negatively correlated across sectors. The main conclusion of the analysis is that representative agent, perfectly competitive economies tend to exhibit more substitution between activities than imperfectly competitive economies with heterogeneous agents.

Drawing on this implication, Section IV of the paper looks at the evidence on employment fluctuations over time and across sectors.[6] Our main finding is that employment fluctuations are generally positively correlated across sectors at the frequency of monthly data. Further, we find that these fluctuations are not due solely to common shocks. We argue that this provides support for the underemployment equilibrium model relative to a representative agent, real business cycle model in which shocks induce the reallocation of workers across sectors.

I. Imperfect Competition, Propagation and Inventories

To study these issues, consider a two-period imperfectly competitive economy. There are two goods produced by monopolists and a third, nonproduced good, endowed to a group of agents termed *outsiders*. There are also workers present each period who sell their labor services to firms.

The monopolist in each of the two sectors produces a unit of output from each unit of labor. In making production decisions, firms recognize their market power as sellers. These agents consume the nonproduced good in each period and, as buyers, act as price takers. Their preferences (formally those of their shareholders) are given by $V(m_t)$ in each period where $V(\cdot)$ is increasing and concave and m_t is their consumption of the nonproduced good in period t. To produce, these firms hire workers in a competitive labor market.[7]

The sectors differ in two important respects. First, in period 1 the monopolist in sector 1 receives an endowment of that good, e_1. Second, this agent can hold goods over time as inventories. The sector 2 monopolist has a zero endowment and is unable to hold inventories. This specification permits us to analyze the interaction between sectors holding inventories (for example, manufacturing) and those for which inventory holding is impossible (for example, services).

There are two types of outsiders. Type i outsiders have an aggregate endowment of M^i units of the nonproduced good in *each* of the two periods.[8] They demand both the nonproduced good and the good produced in sector i and have homothetic preferences. Their demands for sector i output and the nonproduced good are given by

$$q^i = h(p^i)M^i \tag{1}$$

and

$$m^i = M^i(1 - p^i h(p^i)).$$

With homothetic preferences, demands are proportional to income so that variations in M^i induce proportional variations in the demand for the produced good. The function $h(\cdot)$ is decreasing and, since demands are positive, $1 > ph(p)$ for $p > 0$.[9]

---

[6] This empirical exercise follows Long-Plosser (1987) though our emphasis is on employment fluctuations while their work concentrates on output. Further, they focus on manufacturing at the two-digit level while our sectors are defined at the one-digit level. This permits us to include in our analysis the service sector for which inventory holdings are presumably not very important.

[7] This assumption of competitive labor markets is reasonable if the economy was composed of a large number of monopolized sectors. In our two-sector model, this assumption is a bit strong. Cooper (1986) analyzes an alternative way of characterizing labor market equilibrium through a contracting process in which the variations in the market power of workers and firms are easily analyzed.

[8] A natural candidate for the nonproduced good is land, see Hart (1982). As our model is specified, the nonproduced good could be endowed to the workers instead of the outsiders. This nonproduced good generates an autonomous consumption demand for the produced goods.

There are $2N$ workers born each period who are endowed with a unit of leisure. Workers are of two possible types depending on the location of their employment. Those employed in sector i consume the nonproduced good and the good produced in the *other* sector (denoted hereafter by $-i$). This is a common preference structure for these models (see, for example, Cooper and Andrew John, 1988, or Walter Heller, 1986) as it highlights the fact that agents generally consume goods other than those which they produce.

For simplicity, workers in sector $-i$ have the same preferences as type $-i$ outsiders. These workers consume $h(p^i)$wn units of good i and $(1-h(p^i)p^i)$wn units of the nonproduced good where w is the wage rate and n denotes the hours worked. Their indirect utility from selling their entire unit of leisure time at wage w and buying sector i output at price p^i is given by $Z(w,p^i) \equiv U(h(p^i)w,(1-h(p^i)p^i)w)) - k$, where $k > 0$ is the disutility of working and $U(\cdot)$ is increasing, concave and homogeneous of degree one. As a consequence, the indirect utility from working $n < 1$ units of time is simply $nZ(w,p^i)$. $Z \geq 0$ is necessary for worker participation in the market.[10]

Before proceeding it may be fruitful to evaluate the assumptions made thus far. We have restricted attention to two periods and two sectors for ease of presentation. Extensions to more time periods and sectors would not dramatically affect our results.

One important assumption is that there are monopolists in the two sectors. If these markets were oligopolistic, the analysis would be much more difficult as it would require the characterization of a dynamic Cournot-Nash equilibrium with inventories. In this setting, inventories would play an important strategic role as discussed by Lanny Arvan (1985) and Julio Rotemberg and Garth Saloner (1989). While this is an interesting possibility, it did not seem crucial to our line of inquiry and is avoided by the monopolistic setting.

Finally, the structure of demands here is quite important. The monopolists are assumed to consume only the nonproduced good while the workers consume the produced goods.[11] This implies that the current spending of workers is the key linkage across the sectors of the economy and firms' profits represent a leakage from the expenditure stream. As a consequence, changes in the level of production in sector 1 will influence the demand for sector 2 goods. That is, the linkage across sectors is through production *not* sales.

The theme that increased economic activity in one sector creates an incentive for increased activity in the other sector is common to models of coordination stemming from intersectoral spillovers as emphasized in Cooper-John (1988). The structure of demands assumed here is but one of many approaches to generating these spillovers.[12]

To characterize the equilibrium for this two-period model, we first consider the equilibrium in period 2 taking as given the inventory holdings of the sector 1 monopolist. We then consider the first-period choice problems and the overall equilibrium.

Period 2 Analysis for Given Inventories. In period 2, the sector 1 firm has inventories from the previous period which we denote by $I_1 \geq 0$. This monopolist is interested in maximizing income, in terms of its consumption good (the nonproduced good), and solves

---

[9] Additional properties of the demand curves are discussed in Hart (1982). In particular, we will assume that marginal revenue is an increasing function of the price.

[10] If $Z(w,p^i) = 0$, workers will be indifferent with regard to their labor supply. If $Z(w,p^i) > 0$, workers will supply their entire endowment of time.

[11] This is similar to the structure imposed by Hart (1982) in his discussion of the importance of income distribution in models of imperfect competition. The conclusion contains a discussion of a more primitive model which might endogenously generate this structure of demands.

[12] One alternative is to generate these linkages through factor demands as in, for example, Hall (1988) and Cooper-Haltiwanger (1989).

$$(2)\qquad \underset{q_2^1}{\text{maximize}}\ p_2^1(\cdot)\big(q_2^1 + I_1 - w_2 q_2^1.$$

Throughout this presentation, superscripts index the sector and subscripts index the time period. The demand for sector 1 output comes from the outsiders and sector 2 workers and is

$$(3)\qquad q_2^1 + I_1 = h\big(p_2^1\big)\big(M^1 + q_2^2 w_2\big).$$

The inverse demand curve in (2), comes directly from (3). Using (3), the solution to the monopolist's problem is

$$(4)\qquad p_2^1\big[1 + \xi\big(p_2^1\big)\big] = w_2.$$

where $\xi(p) \equiv h(p)/ph'(p) < 0$ is the inverse of the price elasticity of demand.

Because of the homothetic preferences and constant marginal cost, the markup of prices over marginal cost (the wage rate) is independent of the level of output. Thus variations in output and/or inventory holdings will influence prices only when wages are sensitive to employment levels.

The maximization problem of the monopolist in sector 2 is identical to that of the sector 1 monopolist with the restriction that $I_1 = 0$ since there are no inventories in sector 2. The demand for sector 2 output is given by (3) with the appropriate changes in the income level of the consumers of good 2. The solution to the monopolist's optimization problem is therefore

$$(5)\qquad p_2^2\big[1 + \xi\big(p_2^2\big)\big] = w_2.$$

Equations (4) and (5) link the prices in the two sectors to the wage rate in period 2 through $\xi(\cdot)$. Since both firms hire workers at the same wage and face demand curves with the same elasticities, the prices in the two sectors will be equal. Denote by p_2 this common price for period 2.

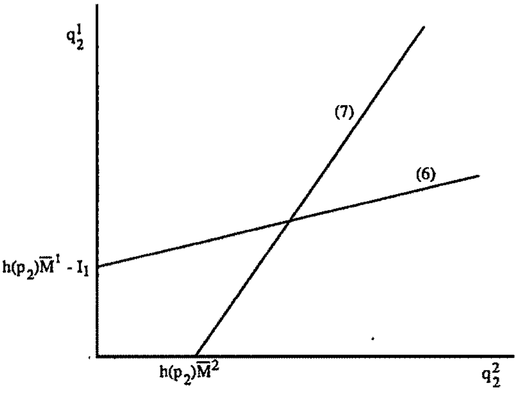


FIGURE 1

Inserting (w_2, p_2) into the sector demand curves yields two equations describing the *output* for each of the two sectors at these prices. These are given by

$$(6)\qquad q_2^1 = h(p_2)\big(M^1 + w_2 q_2^2\big) - I_1$$

for sector 1 and

$$(7)\qquad q_2^2 = h(p_2)\big(M^2 + w_2 q_2^1\big)$$

for sector 2.

For convenience of exposition, (6) is a restatement of (3). As a consequence of homothetic preferences, (6) and (7) imply that output in one sector depends *linearly* on output in the other sector. Note that the demand for sector 2 output depends on the level of output *not* sales in sector 1. This is because the linkage between sectors is through the demands of workers not firms. Hence, a decrease in sector 1 production, perhaps due to an increase in the inventories held by sector 1 firms, will reduce the demand for sector 2 output.

Figure 1 provides a graphical representation of these two curves. Their intersection determines the levels of output for the two sectors for given (w_2, p_2). For I_1 small, both curves have a positive intercept term and slopes less than one, since $p_2 > w_2$ from (5) and $ph(p) < 1$. As the inventory holdings of sector 1 firms increase, the level of production in that sector will fall at each level of sector 2 output, given (w_2, p_2). Since we will ultimately characterize an equilibrium in which wages and prices *will not* vary with I_1, these comparative statics will be important in understanding the linkage between period 1 inventory holdings and the period 2 equilibrium levels of output and employment.

Period 1 *Choice Problems and Equilibrium.* The period 1 choice problem of the monopolist in sector 2, the workers and the outsiders are exactly the same as their period 2 choice problems since none of these agents solve intertemporal problems. In particular, the sector 2 monopolist's first-order condition is given by

$$p_1^2[1+\xi(p_1^2)] = w_1. \tag{9}$$

This is identical to (5) except for the change in the decision period. The sector 1 monopolist selects an output level and inventory holdings to

$$\max V(\pi_1^1) + V(\pi_2^1), \tag{10}$$

where

$$\pi_1^1 = (q_1^1 + e_1 - I_1)p_1^1() - w_1 q_1^1 \quad \text{and}$$

$$\pi_2^1 = (q_2^1 + I_1)p_2^1() - w_2 q_2^1.$$

Here w_2 is the period 2 equilibrium wage anticipated by this firm. The period 2 price function, $p_2^1(\cdot)$ is the inverse demand function that the firm conjectures for the next period. In equilibrium, these conjectures will be correct. In the expression for period 1 profits, e_1 is the monopolist's endowment in the first period. The first-order conditions for this problem are

$$V'(\pi_2^1)MR_2^1 = V'(\pi_1^1)MR_1^1 \quad \text{and} \tag{11}$$

$$p_1^1[1+\xi(p_1^1)] = w_1. \tag{12}$$

In (11), MR_t^1 refers to the marginal revenue of selling an extra unit of sector 1 output in period $t=1,2$. Expression (12) is analogous to (4) and implies that within period 1 marginal revenue will equal marginal cost.

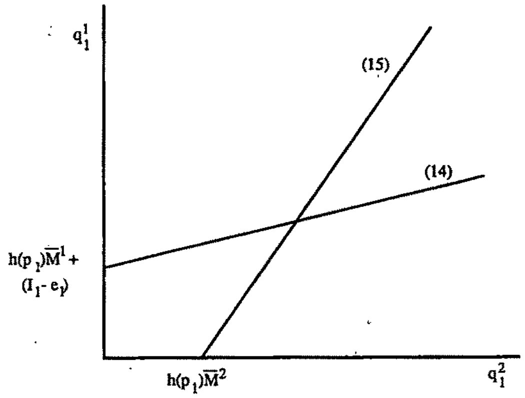


FIGURE 2

Using (12) and (4), (11) can be rewritten as

$$V'(\pi_1^1)w_1 = V'(\pi_2^1)w_2. \tag{13}$$

Expression (13) characterizes the intertemporal decision of the monopolist. The gain to holding more inventories from period 1 to 2 is given by the right side of (13). The monopolist saves the labor cost associated with producing an extra unit times the marginal utility of period 2 income. The cost of holding inventories is the wage for hiring the extra worker times the marginal utility of period 1 income. Since there is no discounting and inventories do not depreciate, $I_1 = e_1/2$. In this economy, as in the Alan Blinder and Stanley Fischer (1981) yeoman farmer model, inventories are used to smooth consumption.

At these wages and prices, the quantities produced in the two sectors are given by expressions analogous to (6) and (7). These are

$$q_1^1 = h(p_1)(M^1 + w_1 q_1^2) + I_1 - e_1 \tag{14}$$

and

$$q_1^2 = h(p_1)(M^2 + w_1 q_1^1). \tag{15}$$

Figure 2 provides a graphical representation of these two equations for a given value of $(I_1 - e_1)$.

Equilibrium. Our interest is in the behavior of the economy in a stationary underemployment equilibrium where total employment and output is the same each period and is strictly less than $2N$. Initially, assume that $e_1 = 0$ and that $M^i = M$ in each period so that preferences, endowments, and technology are all stationary and the sectors are symmetric. Given the symmetry between the sectors of the economy, the equilibrium wages are independent of time and the equilibrium prices are independent of time and sector. We denote these by (w^*, p^*).

In an *underemployment equilibrium*, workers will be indifferent between working and not. Hence the participation constraint

$$Z(w^*, p^*) = 0 \tag{16}$$

will be binding.

To characterize an underemployment equilibrium, solve for the (w^*, p^*) that simultaneously satisfies (16) and one of the identical markup equations, say (4). This is an equilibrium if the labor market clears. Given the stationarity and symmetry, the condition for labor market clearing is

$$q^* = h(p^*)[M + w^*q^*] < N. \tag{17}$$

In a symmetric equilibrium, there will be at most N units of labor time available in each of the two sectors. With q^* units of output demanded in each given (w^*, p^*), (17) is a condition that labor demand does not exceed the labor capacity in the economy. Intuitively, it is a condition on the productive capacity of the economy relative to the demand forthcoming at the prices (w^*, p^*). If (17) holds, workers will be indifferent with regard to the number of hours they supply so that an underemployment equilibrium can obtain.[13]

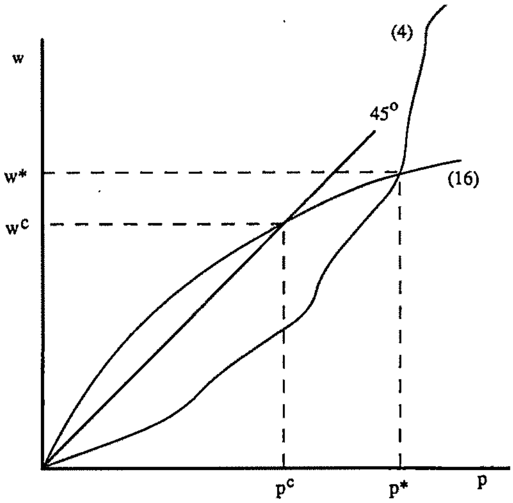


FIGURE 3

PROPOSITION 1: *If* (17) *holds*, $e_1 = 0$ *and* $M^i = M$ *for* $i = 1, 2$, *then there will exist a symmetric, stationary underemployment equilibrium for this economy with prices and wages determined by* (4) *and* (16) *and quantities determined by demand at these prices.*

PROOF:

See Appendix.

Figure 3 depicts an equilibrium (w, p) from the crossing of a worker's indifference curve (16) with one of the markup equations, (4). Note that multiple equilibria are possible if $\xi(p)$ varies enough to create multiple crossings of these curves. This possibility is discussed in Heller (1986). Uniqueness of equilibrium is guaranteed by assuming a constant elasticity specification.

If (17) does not hold, then the economy will be in a full employment equilibrium. In that case, the behavior of the equilibrium prices and quantities in response to shocks is radically different than in the underemployment regime. We emphasize the underemployment equilibria in the analysis to follow and then return to a discussion of the full employment economy in Section III of this paper.

Effects of Changes in Period 1 *Endowment*. Consider the effect on the economy of an increase in the endowment of the monopolist in period 1 at the underemployment equilibrium characterized by Proposition 1. This exercise of varying an endowment is similar to that outlined by Blinder-Fischer (1981) as a vehicle for understanding the *propagation* effects of holding inventories.

---

[13] Similar arguments appear in Cooper (1986), Hart (1982), Hall (1988) and Weitzman (1985). As discussed in Section III, this is a useful structure for understanding the implications of imperfect competition since there are basically two regimes that an equilibrium can fall into (either underemployment or full employment) and changes in market structure influence the likelihood of observing the economy in one or the other regime.

This structure shares some of the weaknesses of Blinder-Fischer (1981) and related models. These models generally predict that inventories play an important role in smoothing production and consumption, a view of the role of inventories that is called into question by evidence suggesting that the variance of production exceeds that of sales. Further, the predictions of these models are not obviously consistent with the observation that inventories are procyclical, as discussed in Blinder (1981). Note, however, that there is some empirical evidence in support of the idea that unplanned inventory accumulation induces a reduction in employment. First, one does typically observe that inventories rise at the initial stages of a slump in demand (although they ultimately decline during slumps). This manifests itself in the finding that there is a statistically significant inverse relationship between initial inventories in a period and the various components of employment (see Haltiwanger and Louis Maccini, 1989). Second, in a similar manner, although inventories are procyclical, the inventory/sales ratio is countercyclical. In fact, innovations in the inventory/sales ratio are good predictors for inflows into unemployment (see Michael Darby, Haltiwanger, and Mark Plant, 1985). Overall, then, our view is that empirical examination of the interaction of inventories and employment requires a careful analysis of both the timing of the interaction and a decomposition of planned (or desired) and unplanned inventory fluctuations.

There are a number of alternative shocks that could be considered in this framework such as fluctuations in marginal products and/or variations in the endowment levels of the outsiders.[14] While this might be of independent interest, for the purpose of focusing on the propagation effects of inventories on future production, consumption and employment, the analysis under either of these two alternatives would be the same as that conducted here.

As long as we stay in an underemployment equilibrium, the (w^*, p^*) given in Figure 3 will be invariant to small changes in e_1. Thus the equilibrium quantities will be given by (6), (7), (13), (14) and (15) evaluated at (w^*, p^*). The equilibrium can be characterized explicitly from these five relations or implicitly through Figures 1 and 2 using (13).

An increase in e_1 has two important effects. From (14), the output of the sector 1 monopolist would fall for given inventory holdings. So, in Figure 2, this causes a shift down in the line expressing sector 1 output as a function of sector 2 output in period 1. As seen from this figure, this causes a reduction in sector 2 output as well. However, from (13), part of this increase in endowment would be held as inventories by the monopolist. From (6), this increased level of inventories will then reduce the sector 1 monopolist's output in period 2. This reduction in sector 1 output in the two periods spills over to sector 2 because sector 1 employment falls. A decrease in sector 1 employment reduces demand for sector 2 output since the linkage across the sectors is based on the demands of the workers. This is seen in Figure 1 as a shift down in the line expressing sector 1 output as a function of sector 2 output in period 2. These effects are summarized by

PROPOSITION 2: *An increase in* e_1 *will increase* I_1 *and reduce* q_t^i *for* $i = 1,2$ *and* $t = 1,2$.

PROOF:

From (11), if an increase in e_1 increases π_1^1, then I_1 must increase so that this condition is met. However, the increase in I_1 will not exceed the increase in e_1. Holding inventories constant and using (12), an increase in e_1 increases π_1^1 by MR_1^1. So, in order for (11) to hold I_1 must increase with e_1 but by less than e_1.

---

[14] As a referee pointed out, it is straightforward to take the conditions describing an underemployment equilibrium and totally differentiate the system of equations with respect to the endowment of the outsiders and to introduce a technology shock.

From (14), this implies that period 1 output for sector 1 will fall and from (6) period 2 output will fall as well. Since sector 2 output is an increasing function of sector 1 output (not sales) in both periods, sector 2 output will fall as well. ■

The key to this result is the positive interaction across sectors in the economy. Reduction in sector 1 output in each of the periods is a consequence of a positive endowment shock and the smoothing behavior of inventory holdings as the monopolist seeks to spread the endowment increase over time. The resulting reduction in sector 1 output spreads to sector 2 through the reduced demands of the unemployed workers. Thus these demand spillover effects, coupled with the persistence created by the holding of inventories, generates correlated movements in output and employment across sectors and time.

II. A Representative Agent Model with Sectoral Reallocation

In this section we consider an alternative approach to studying the propagating effects of inventories. In particular, we construct a representative agent, real business cycle model in which one of the two commodities can be held in inventory. Given the success of this approach in matching some features of the business cycle, see, for example, Finn Kydland and Edward Prescott (1982), Long-Plosser (1983) and Gary Hansen (1985), it is of interest to understand the implications of this approach for the comovements in output and employment when inventories provide an important intertemporal link.[15]

The model differs from that in Section I both in terms of the market structure and in the specification of agents' preferences. In particular, we follow the real business cycle literature by assuming that the competitive market allocation corresponds to the solution of a planner's problem and that there is a single agent in the economy. These differences in assumptions and their implications for our results are discussed in some detail in Section III of the paper.

Consider a Robinson Crusoe economy in which the single agent has preferences described by

$$(18)\quad \sum_{t=1}^{2} \beta^t \{ U^1(c_t^1) + U^2(c_t^2) + U^3(m_t) - g(n_t) \},$$
$$n_t = n_t^1 + n_t^2,$$

where $U^i(\cdot)$ is increasing and concave, $g(\cdot)$ is increasing and convex and both are twice continuously differentiable. As in the Section I model, there are two produced goods (c_t^1, c_t^2) and one nonproduced good (m_t) in each period. Robinson Crusoe is endowed with $2N$ units of leisure and M units of the nonproduced good in each period. The production technology is given by

$$(19)\quad q_t^i = f^i(n_t^i) \quad i = 1,2,$$

where $f' > 0$ and $f'' < 0$. Good 1 is storable while good 2 is not. At the beginning of period 1, Robinson Crusoe receives an endowment of good 1, e_1.

Robinson Crusoe maximizes utility given by (18) subject to the technology given by (19) and the following feasibility constraints:

$$(20)\quad c_t^1 + I_t = q_t^1 + I_{t-1} \qquad \text{for } t = 1,2,$$

$$(21)\quad c_t^2 = q_t^2 \qquad \text{and}$$

$$(22)\quad m_t = M,$$

where $I_t \geq 0$ is end-of-period inventories and $I_0 = e_1$. As there is no incentive to hold inventories at the end of period 2, $I_2 = 0$. Given this, the optimality conditions (for an interior solution to n_t^1, n_t^2, and I_1) reduce to

---

[15] Hence we diverge from Long-Plosser (1983) and focus on the linkages over time created by inventories and ignore for now the possibility that commodities of one type might be important elements in the production of other commodities. The model is close to that explored by Rogerson (1987) though we focus on the role of inventories and abstract from mobility costs.

$$(23)\quad u^{i\prime}(c_t^i)f^{i\prime}(n_t^i)=g'(n_t);\qquad i=1,2;\ t=1,2\ \text{and}$$

$$(24)\qquad U^{1\prime}(c_1^1)=\beta U^{1\prime}(c_2^1).$$

Expression (23) implies that Robinson Crusoe makes his within-period consumption and production decisions so that the marginal utility from the output produced by one additional unit of labor equals the marginal disutility of an additional unit of labor. Expression (24) indicates that inventories are used to smooth the marginal utility of consumption over time.

Now consider the same conceptual experiment conducted in Section I. In particular, consider the effect on the economy of an increase in the endowment of good 1 in period 1. As the following proposition demonstrates, in response to an increase in e_1, Robinson Crusoe will reduce his production in sector 1 and *shift resources* toward the production in sector 2 in both periods.

PROPOSITION 3: *An increase in e_1 will increase I_1, decrease q_t^1, and increase q_t^2 for $t=1,2$.*

PROOF:

As long as $e_1+q_1^1$ increases in response to the increase in e_1, then both c_1^1 and I_1 must increase in order to satisfy (24). Further, $e_1+q_1^1$ not increasing would violate (23). Observe from (23) that the increased consumption of good 1 in period 1 will induce a reduction in the production of good 1, which in turn by reducing the marginal disutility of labor induces an increase in the production and hence consumption of good 2. Further, since I_1 increases as well this will induce an increase in the consumption of good 1 in period 2. This, in turn, will induce a similar reduction in the production of good 1 in period 2 but an increase in the production of good 2 in period 2. ■

Overall, we observe that in this setting the comovements in inventories, output and employment are very different than those found in the model of Section I. In this environment with complete mobility across sectors and resources necessarily being allocated to their highest value use, inventory movements in one sector induce *substitution* of resources away from that sector to the other sectors of the economy. Thus neither output nor employment vary together in response to endowment shocks. The negative comovements generated in this model are similar to those reported in the sectoral shifts literature (for example, Richard Rogerson, 1987) and contrasts with the positive comovement in output and employment present in the specification of Section I. In the subsequent section we discuss what factors underlie the dramatic differences in the predicted behavior of inventories, output and employment across the two alternative specifications.

Before proceeding, it is useful to bridge the conceptual gap between Propositions 2 and 3 and the results that emerge from the Long-Plosser analysis. In contrast to Proposition 3, Long-Plosser demonstrate that their model can produce correlated movements in output across sectors which reflect technological complementarities characterized through an input-output matrix. As noted earlier, their explicit example yields an allocation in which employment levels in each sector were constant because of the Cobb-Douglas preferences.

Their model differs from ours in two important ways. First, the shock in their model is sector-specific productivity disturbance which occurs *after* the determination of factor inputs. That is, their model contains a stochastic technology with a production lag while we focus on an endowment shock. Second, the use of commodities as inputs in the production of other commodities provides a mechanism for generating positive comovements in production across sectors and time. In the Long-Plosser model, these linkages are characterized by an input-output matrix with nonzero off-diagonal elements. Our model has a matrix in which the only linkage across commodities and time is the holding of sector 1 inventories.

To understand the differences between these models, consider a transitory productivity shock in sector 1 in the model of this section instead of an endowment shock while retaining the assumption of no production lags. The optimality conditions for the planner's problem will equate the marginal disutility of work with the product of the marginal productivity of labor in a given sector and the marginal utility of consuming that sector's output (as in (23)). Suppose that there is a positive productivity shock in sector 1. If $U^1(c_t^1)$ is not too concave, then employment and output in sector 1 will increase while sector 2 output and employment will decline.[16] Through inventory holdings, the effects of the productivity shock will be propagated over time. Thus the response to a shock to the sector 1 production function would be the substitution of employment across the sectors just as in Proposition 3 so that *negative* correlations in output and employment across sectors would again be predicted.

To move a step closer to the Long-Plosser specification, suppose that sector 1 output is produced with a lag so that period t output in sector 1 is a function of labor input in period $t-1$ and a shock in period t. Further, assume that sector 2 does not have a production lag (i.e., continue to think of that as a service sector). Then, in an infinite horizon version of the model (which is needed to avoid the inconsistency of having a final period in a model with a production lag), a shock to the sector 1 production function in period t will reduce employment in that sector and cause the substitution of labor into sector 2. Output will be higher in both sectors in the period of the shock though sector 1 output will, on average be lower in the next period (assuming i.i.d. shocks) due to the reduction of sector 1 labor input. Note that once again, employment is negatively correlated across sectors as the productivity shock creates substitution effects.

As a final step, suppose that we allow the production of commodities by commodities and assume that labor and other inputs are complements in the production function, as in the Long-Plosser model. A shock to the sector 1 production function will then cause a decline in the marginal utility of consuming sector 1 output. Since some of this additional output is used in the production of sector 1 goods *next* period, this increases the marginal productivity of labor. Thus, as discussed by Long-Plosser, there is an ambiguity regarding the response of employment to productivity shocks in the Long-Plosser setting. A sufficiently large technological complementarity between inputs and labor in *all* sectors relative to the curvature of the utility function is required to generate positively correlated movements in employment. In a sense, the effect of the technological linkages between sectors in the Long-Plosser model can be similar to an economywide technology shock inducing both income and substitution effects and hence an ambiguous effect on employment in each sector. If, however, there are sectors in the economy for which the marginal product of labor is *not* an increasing function of the level of intermediate inputs, then employment in these sectors will be *negatively* correlated with employment in sectors exhibiting the technological complementarities.

The point of this analysis is to demonstrate the importance of the reallocation of labor in the presence of sector specific technology shocks. This effect will generally produce negative correlations in sectoral employment levels.

Together, Propositions 2 and 3 illustrate two important forces which determine the effects of shocks, propagated through inventory holdings, on output and employment over time and sectors of the economy. On the one hand, shocks to one sector which reduce the level of activity in that sector spillover to other sectors through demand linkages. These affects can be propagated over time through inventory holdings in at least one of the sectors. On the other hand, reductions in activities in one sector allow for the substitution of resources to other sectors. This is the point of Proposition 3. From the standpoint of evaluating potential models of fluctuations, it is important to understand the characteristics of economies which give rise to these alternative implications and to evaluate these effects empirically. This is the point of the next two sections of the paper.

---

[16] Even if $U^i(\cdot)$ is so concave that income effects dominate, as long as $g(\cdot)$ is strictly convex, productivity shocks still generate substitution of employment across sectors.

III. Reconciliation

There are several factors that contribute to the differences in predicted behavior across the specifications presented in Section I and II. One important element is the imperfect competition of the Section I model. To see this, consider the equilibrium allocation from a competitive version of that model. An underemployment equilibrium may still arise and is characterized by simultaneously solving the participation equation (16) and the zero profit condition of $w_t = p_t^i$ for all t and i. The competitive stationary, underemployment solution is given by the (w, p) solving these two equations *if* the resulting quantities demanded are less than N in each sector. If they are not, then a full employment equilibrium results.

A competitive, underemployment equilibrium (w^c, p^c) is depicted in Figure 3 by the intersection of (16) with the 45° line. Since $\xi(p) < 0$, the wage and price levels are *higher* for the imperfectly competitive economy than under perfect competition in an underemployment equilibrium. This is indicated in Figure 3 by the fact that the curve for the markup equation in the imperfectly competitive economy lies below the 45° line. As a consequence, when $\varepsilon_1 = 0$, the stationary competitive level of output (q^c) is greater than q^*. This is indicated in Figure 4 which depicts the analogue of Figures 1 and 2 for the stationary underemployment equilibria in the two market structures when $\varepsilon_1 = 0$.[17]

The set of parameter values for which an underemployment equilibrium can occur (q^c less than N) with perfect competition is a subset of that for imperfect competition.[18] That is, if the parameters describing the economy are chosen randomly at the start of the model, then imperfect competition increases the probability that the stationary equilibrium will be in the underemployment region. As a consequence, the propagation effects of inventories on output and employment discussed above are more likely to occur. Conditional on being in the underemployment region, the quantitative impact of the positive spillover effect is actually *greater* under perfect competition. This can be understood by noting in Figure 4 that the slopes of the two curves associated with perfect competition are greater than those associated with imperfect competition.[19]

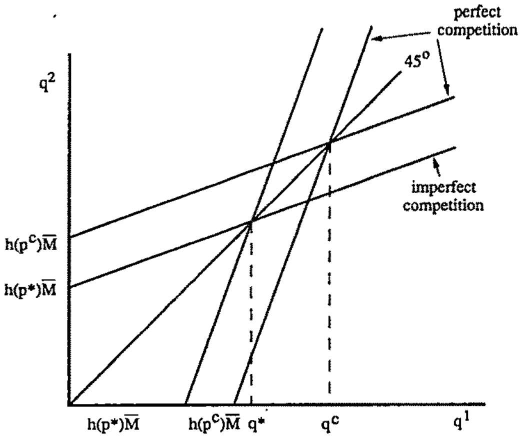


FIGURE 4

---

[17] When $\varepsilon_1 > 0$, for an underemployment equilibrium production is no longer symmetric across sectors. However, the same arguments used here will imply that production in each sector under imperfect competition will be lower than that under perfect competition.

[18] The relationship between the size of the underemployment region and the market structure depends on the elasticity of demand through the markup and on the relative budget shares of the produced and nonproduced goods in both the competitive and imperfectly competitive allocations. Without further assumptions on functional forms, the relative size of these regions is difficult to quantify. From Figure 3, one can see that if the markup is increased, the curve labeled (4) shifts to the right leading to higher prices and lower output. In a related model, Hall (1988) shows that if the markup is about 1.5, then the region of underemployment equilibria under imperfect competition is 1.5 times that of the perfectly competitive economy. While these calculations are not directly applicable to our model, they do give some indication of the relative sizes of these regions in a related model.

[19] The slope of each curve under perfect competition is given by $h(p^c)p^c$ while the slopes under imperfect competition are given by $h(p^*)w^*$. It can be easily shown that the slope under perfect competition is larger.

Three remarks are in order regarding this issue. First, under both market structures, the correlation coefficient of employment across sectors equals one, conditional on being in the underemployment region. Second, since the probability of being in an underemployment equilibrium is greater under imperfect competition, the observed correlation coefficient of employment movements across sectors will be *higher* under imperfect competition. Finally, as our discussion emphasizes, imperfect competition is a contributing factor to the dramatically different predictions from the models of Sections II and III but it is not the only factor.

These arguments are further strengthened by noting that there are environments in which an underemployment equilibrium occurs *only if* some agents have market power. Hart (1982) provides such an example in which the market power of workers in the labor market supports an underemployment equilibrium in an economy for which all perfectly competitive equilibria are full employment solutions. Heller (1987) describes an economy in which workers have a zero value for leisure so that all competitive equilibria are full employment allocations. In that setting, Heller describes the possibility of an underemployment solution. In our model, setting $k = 0$ would sever the linkage between sectors since wages would be driven to zero and workers would have no income to spend on the other sector. This problem can be circumvented either by allowing firms to spend some of their income in the other sector or by considering a contracting solution in the labor markets in which workers have some bargaining power.[20]

The point of the emphasis on underemployment equilibria is that the results reported in Proposition 2 only hold in the underemployment region. In the event that the economy is at a full employment solution, we obtain results very similar to the representative agent model of Section II. That is, an increase in initial inventories in a particular sector induces a reduction in the production in that sector and a shift of resources toward the other sector. Thus the empirical observation that employment is positively correlated across sectors over the cycle (see the discussion in the next section) is produced by an economy operating in the underemployment region.

Another important part of the Section I model is the extreme nature of the disutility of work. The fact that k is common to all workers creates an inverted L-shaped labor supply curve and hence the two regimes noted above. As a consequence, the behavior of the economy across the two regimes is vastly different.

The models in Sections I and II do differ with regards to the specification of the disutility of work and these differences deserve comment. If we replaced kn in the Section I model with $g(n)$, where $g(n)$ is increasing and strictly convex as in the Section II model, then wages and prices would vary with output and employment and this would reduce the impact on production and employment of the demand linkages across the sectors.[21] Alternatively, if $g(n_t) = kn_t$ in the Section II model, then the connection between sectors of the economy would be served. As a consequence, shocks to one sector would have no effect on output and employment in the other sector.

Finally, the model explored in Section I contains strong assumptions about the heterogeneous preferences of workers and firms. As discussed earlier, this heterogeneity implies that consumption of produced goods is sensitive to the current income of workers. This structure creates a linkage between the production in one sector and the output level of the other sector which is absent from the representative agent model of Section II. In the latter model, consumers are effectively aggregated and their preferences over produced goods are represented by the functions $U^i(c_t^i)$ for $i=1,2$ and $t=1,2$.

---

[20] This approach is described in Cooper (1986).

[21] As long as marginal cost is convex in output, the variation in quantities relative to prices will be much larger when firms have market power. This is because the market power of firms will yield higher prices and lower output and employment. As long as marginal cost is convex, variations in marginal revenue will then produce more fluctuations in quantities and less in prices as a consequence of the market power. The point of the extreme specification adopted here is to highlight the fact that imperfect competition implies that the economy is more likely to be operating on a flat portion of the aggregate labor supply curve.

To see the importance of the differences in tastes, suppose that in the Section I model, workers and firms had identical homothetic preferences. Then, total demand for each good would simply depend on aggregate income as if there was only a single consumer. Spending on each sector would then depend *only on* the revenues generated by production in the other sector and not the decomposition of these revenues into profits and wages. As a consequence, the optimal sales of the sector 1 firm would be independent of the endowment and inventory holding of that firm and the level of activity in sector 2 would be independent of these variables as well. Thus, fluctuations in sector 1 inventories caused by variations in e_1, would *not* spillover to activities in sector 2.

Overall, the models in Sections I and II capture some important forces in the effects of sectoral shocks on employment at the disaggregated level. The specification of preferences coupled with the operation of the economy in an underemployment region produces positive comovements in the model of Section I. Sector-specific shocks, however, may cause the substitution of workers from one sector to another as captured by the model of Section II. While these sectoral reallocation effects may be reduced by mobility costs, as in Rogerson (1987), they tend to produce negatively correlated employment fluctuations.

IV. Empirical Implications

The models explored in Sections I and II have interesting empirical implications regarding fluctuations in output, employment and inventories both over the cycle and across sectors. In particular, the analysis suggests that models with imperfect competition may generate time series with different properties than perfectly competitive economies. To emphasize an important message, imperfectly competitive economies are, on average, predicted to exhibit less substitution across activities than their perfectly competitive counterparts.

One difficulty with confronting these models with data concerns the source of fluctuations in the economy. In most of our formal models we rely on endowment shocks to the inventoriable good as the initial source of fluctuations. Broadly speaking, these disturbances can be interpreted as resulting from sector-specific shocks to tastes or technology. However, we have not formally examined whether the models of perfect and imperfect competition respond differently to aggregate demand and technology shocks.[22] We conjecture though, that once inventories holdings respond to temporary shocks, our results that imperfectly competitive economies tend to exhibit positive correlations in employment and output in subsequent periods will remain.

An important implication of our work is that inventory holdings in some sectors of the economy create correlated behavior in output over time in all sectors of the economy. This is a consequence of the demand spillovers across sectors in a given time period. Thus one need not appeal to inventory holdings in all sectors of the economy as an explanation for observed correlated behavior economywide.

Another important implication of our analysis concerns the comovement of output and employment across sectors of the economy. The analysis in Section I suggests that through the markup of price over marginal cost, imperfectly competitive economies are more likely to exhibit positively correlated fluctuations in output and employment. In contrast, the models explored in Section II are more likely to imply that employment fluctuations are *negatively* correlated across sectors. Shocks to production functions in one sector of the economy cause a reallocation of labor services from one sector to another.

---

[22] The discussion at the end of Section II provides a preliminary extension of Proposition 3 to technology shocks. See Cooper (1986) for a discussion of demand and supply disturbances in a model of imperfect competition.

A complete analysis of the empirical predictions of the models in Sections I and II is beyond the scope of this paper. However, a brief examination of the data on employment across sectors is instructive to identify relevant empirical regularities. In this regard, we examine the time-series and cross-section behavior of employment across sectors. In the subsequent empirical analysis, we follow the methodology used by Long and Plosser (1987) very closely. There are, however, a few important differences that motivate our independent empirical investigation. First, they examined output movements as opposed to employment. Our examination of employment is important in light of Propositions 2 and 3. Second, they examine the output movements for the 13 commodity output components of the industrial production index. While this is appropriate for their examination of the role of technological complementarities, it is less so for our examination of the role of final goods demand linkages particularly between inventory and non-inventory holding sectors.

The data we use to measure employment are monthly observations on seasonally adjusted total manhours by sector for 1947:1 to 1985:12. The methodology we use is as follows. First, we examine contemporaneous correlations of detrended employment (using linear and quadratic time trends) in order to evaluate the overall nature of comovements in employment across sectors.[23] Second, we attempt to evaluate the role of common shocks in generating the observed comovements. To accomplish this, we follow Long and Plosser by decomposing employment movements in period t into functions of shocks (or innovations) that occur in period t and decisions based upon information available prior to period t. If common shocks explain the observed comovements of cyclical employment, then the correlation matrix of contemporaneous employment shocks will have large off-diagonal elements and will be consistent with a small number of unobservable common factors generated from factor analysis. Essentially, this entails generating the contemporaneous correlation matrix of employment innovations and comparing it with the contemporaneous correlation matrix of overall employment movements. It is important to note in this regard that the comparison is between correlation matrices (not covariance matrices) so that comparison of magnitudes of the off-diagonal elements is appropriate.

Again following Long and Plosser, to identify empirically the monthly employment innovations we use the estimated residuals from a vector autoregressive specification. Specifically, we estimated a VAR with the eight employment series using log values, a lag length of six and including linear and quadratic time trends in the estimation.

In Table 1, the lower off-diagonal elements are the correlations of the detrended employment levels while the upper off-diagonal elements are the correlations of the employment innovations. Several features are worth noting. First, virtually all of the lower off-diagonal terms are positive and relatively large in magnitude. This indicates that employment exhibits substantial positive comovement across sectors over the cycle. Further, the inventory holding sectors (i.e., manufacturing and wholesale and retail trade) are quite important in this regard. That is, employment in the inventory holding sectors is highly and positively correlated with employment in the non-inventory holding sectors (for example, services) over the cycle.

In contrast, the upper off-diagonal elements are relatively small. A comparison between the lower and upper off-diagonal elements suggests that the observed positive comovements are not primarily due to common shocks. For example, the lower diagonal correlation in Table 1 between manufacturing and services is 0.63 but the corresponding upper diagonal correlation is 0.08. This suggests that in this case common

---

[23] We also considered alternative detrending methods, such as log first differences. Results on correlations in employment changes across sectors are similar to those reported in Table 1 and the results of the factor analysis using log first differences are similar to those reported in Table 2. However, this approach did seem to overdifference the data and hence generate negative serial correlation.

TABLE 1—CONTEMPORANEOUS CORRELATIONS OF EMPLOYMENT AND EMPLOYMENT INNOVATIONS[a]

	CON[b]	FIR	GOV	MFG	MIN	TRC	SER	WRT
CON	1.00	0.06	0.30	0.30	0.13	0.09	0.12	0.28
FIR	0.62	1.00	0.05	0.06	−0.05	0.00	0.10	0.15
GOV	0.19	−0.04	1.00	0.17	0.10	0.10	−0.01	0.06
MFG	0.61	0.29	0.54	1.00	0.25	0.27	0.08	0.31
MIN	0.23	0.31	−0.13	0.14	1.00	0.26	0.04	0.11
TRC	0.71	0.44	0.56	0.76	0.13	1.00	−0.05	0.11
SER	0.22	0.20	0.51	0.63	0.06	0.39	1.00	0.16
WRT	0.60	0.51	0.59	0.79	0.28	0.76	0.58	1.00

Data Source: Employment and Earnings, Bureau of Labor Statistics, Table C-9.

[a] Using monthly seasonally adjusted data, 1947 to 1985. Lower off-diagonal elements are correlations of log, detrended employment. Upper off-diagonal elements are correlations of innovations to log, detrended employment—see text for further explanation.

[b] Abbreviations: CON, Construction; FIR, Finance, Insurance, Real Estate; GOV, Government; MFG, Manufacturing; MIN, Mining; TRC, Transportation and Public Utilities; SER, Services; WRT, Wholesale and Retail Trade.

shocks account for less than one-seventh of the observed positive comovement in manufacturing and service employment. Other comparisons tend to yield similar conclusions, although there are a few exceptions.

Note that the upper off-diagonal elements of Table 1 may reflect the propagation of sectoral shocks as well to the extent that this propagation is contemporaneous. Hence, the small upper off-diagonal elements of Table 1 may be interpreted as evidence against both the common shock view and contemporaneous propagation of sectoral shocks. It should be noted in this regard that our formal model of Section I predicts contemporaneous propagation of sectoral shocks (as well as lagged propagation through the role of inventories). However, the formal introduction of lags in the cross-sectoral final demand linkages would alter the precise predictions in a manner consistent with the finding of a lack of contemporaneous propagation at high frequencies.[24]

In order to gain further insight into the role of common shocks, we again follow Long-Plosser (1987) and perform a simple factor analysis on the innovations. Since factor analysis decomposes the variation in a set of variables into that due to unobserved common factors and that due to idiosyncratic disturbances, we can obtain some measure of the role of common shocks in explaining the innovations. Note that since the estimated common factors are not constrained to correspond to any observable aggregate disturbance this method overstates the role of any specific aggregate disturbance. In Table 2, we report the fractions of innovation variance explained by common factors for both a one- and a two-factor model. For a one-factor model, the proportion of variance explained ranges from a low of 0.02 to a high of 0.33. The addition of a second factor adds some additional explanatory power but mostly for mining, transportation and services. Overall, these numbers indicate that some but clearly not all of the variation in the innovations can be ex-

---

[24] Suppose, for example, that consumption by workers is based upon last period's income (either due to credit or cash-in-advance constraints). Then the propagation of sectoral shocks across sectors due to final good demand linkages would not be contemporaneous. We did not explicitly model the final goods demand linkages in this fashion for reasons of analytical simplicity. Our primary objective in Section I was to demonstrate through the use of a relatively simple general equilibrium model that complementarities generated by final goods demand linkages may generate positive comovement both across sectors and across time. In this regard, we formally incorporated only one intertemporal linkage (i.e., inventories) but clearly other intertemporal linkages (for example, those that induce consumption lags) may be at work as well.

TABLE 2—FACTOR DECOMPOSITION OF INNOVATION VARIANCE[a]

	Factor Models[b]	
	$R^2 1F$	$R^2 2F$
CON	0.26	0.28
FIR	0.02	0.06
GOV	0.11	0.11
MFG	0.33	0.33
MIN	0.13	0.19
TRC	0.12	0.21
SER	0.03	0.09
WRT	0.21	0.24

[a] See notes from Table 1 for variable definitions and data source.

[b] 1- and 2-factor models of innovations (i.e., residuals from the VAR). The R^2 are fractions of innovation variance explained by the common factors.

plained by a small number of common factors.

One question in interpreting our results is whether the use of monthly data is appropriate. If common shocks do not impact all sectors at exactly the same time, then measuring the correlation of contemporaneous innovations may be inappropriate. This is a difficult issue and a satisfactory solution is beyond the scope of this paper. One simple means of addressing this issue is to examine the results using lower frequency data. Examination of lower frequency data presumably lessens the likelihood that there is a problem due to common shocks not impacting all sectors in the same period. On the other hand, examination of lower frequency data generates new interpretation problems. This is because the lower the frequency of the data the greater the likelihood that a high observed contemporaneous correlation between innovations represents the propagation of a sectoral shock.

To shed some light on the sensitivity of the results to the frequency of the data, we recalculated the correlations using quarterly data. These results are reported in Table 3. As with the monthly data, the lower off-diagonal elements reveal substantial positive comovement across sectors. Compared to Table 1, the lower off-diagonal elements are slightly larger. As expected, the upper off-diagonal elements in Table 3 tend to be larger than the equivalent elements of Table 1. Nevertheless, most of the upper off-diagonal elements of Table 3 are still less than half of the corresponding lower off-diagonal elements. This suggests that even with the use of quarterly data much of the observed positive comovement in employment across sectors cannot be attributed to common shocks.

Another key issue is the persistence in the employment movements across the sectors. Table 4 presents the first-order own and cross correlations across the sectors. Observe that the detrended employment series all exhibit a high and similar degree of positive serial correlation. Further, the cross lagged correlations are mostly positive and often large in magnitude. Note in particular that lagged values of employment in the inventory holding sectors are highly correlated with contemporaneous values of employment in all other sectors.[25] This suggests that changes in employment in the inventory holding sectors often precede changes in employment in the non-inventory holding sectors.

While these results are only suggestive, they are consistent with the predicted patterns of Section I but are not consistent with the predicted patterns of Section II. That is, there are positive comovements in employment fluctuations across sectors and across time. The inventory holding sectors seem to play a fundamental role in this regard as changes in employment in the inventory holding sectors precede and contemporaneously move with changes in employment in the non-inventory sectors. Further, these patterns do not appear to be driven by the presence of common shocks. Qualitatively, Long-Plosser (1987) report similar findings in their investigation of output comovements.

---

[25] In particular, the average correlation between lagged employment in the manufacturing sector and current employment in each sector is 0.60 while the analogous average correlation using lagged wholesale and retail trade is 0.62.

TABLE 3—CONTEMPORANEOUS CORRELATIONS OF EMPLOYMENT AND EMPLOYMENT INNOVATIONS[a]

	CON	FIR	GOV	MFG	MIN	TRC	SER	WRT
CON	1.00	0.24	0.34	0.32	0.25	0.21	0.69	0.39
FIR	0.65	1.00	0.00	0.26	0.13	0.08	0.22	0.28
GOV	0.21	−0.05	1.00	0.09	0.14	0.13	0.06	0.00
MFG	0.65	0.31	0.55	1.00	0.24	0.57	0.19	0.51
MIN	0.25	0.34	−0.16	0.14	1.00	0.36	0.13	0.25
TRC	0.73	0.44	0.60	0.78	0.14	1.00	0.09	0.36
SER	0.27	0.23	0.53	0.65	0.08	0.44	1.00	0.23
WRT	0.65	0.54	0.20	0.81	0.29	0.80	0.62	1.00

[a] Using quarterly data from 1947 to 1985. Lower off-diagonal elements are correlations of log, detrended employment. Upper off-diagonal elements are correlations of innovations to log, detrended employment. See Table 1 for variable definitions and data source.

TABLE 4—CORRELATION MATRIX OF THE LAGS OF EMPLOYMENT AT LAG 1[a]

	CON	FIR	GOV	MFG	MIN	TRC	SER	WRT
LCON[b]	0.90	0.63	0.17	0.60	0.23	0.70	0.23	0.61
LFIR	0.61	0.95	−0.07	0.27	0.32	0.4s1	0.20	0.50
LGOV	0.16	−0.04	0.91	0.53	−0.13	0.56	0.51	0.59
LMFG	0.60	0.32	0.54	0.99	0.15	0.77	0.64	0.82
LMIN	0.21	0.30	−0.16	0.12	0.86	0.11	0.06	0.26
LTRC	0.70	0.46	0.56	0.75	0.15	0.98	0.40	0.78
LSER	0.19	0.21	0.51	0.60	0.06	0.37	0.89	0.59
LWRT	0.58	0.52	0.58	0.75	0.27	0.75	0.58	0.94

[a] Log, detrended employment. See Table 1 notes for variable definitions and data source.
[b] The prefix L indicates the lagged series.

V. Conclusion

The twin observations of serially correlated output movements and positive comovements in output across sectors of a multisector economy are two important "stylized facts" that guide macroeconomic researchers. The important paper by Long-Plosser (1983) provides one possible explanation for these features based on the normality of consumption goods in preferences and a technology which allows for the production of commodities by commodities. As a consequence, the linkage highlighted in their paper are intersectoral demands for intermediate goods.

This paper investigates an alternative approach which stresses linkages across sectors from the demand for consumption goods and the presence of imperfect competition. The market power of sellers produces an environment in which an underemployment equilibrium is likely to arise. In that case, shocks to one sector of the economy spillover to others and generate positively correlated output and employment movements. The linkage across time explored in this paper is the holding of inventories by a *subset* of agents in the economy. Because of the demand linkages across sectors, the holding of inventories by some firms is enough to produce correlated output movements in all sectors.

This model is then contrasted with a representative agent, real business cycle model *without* the factor demand flows stressed by Long-Plosser. For that setting, positively correlated employment fluctuations across sectors are not likely to occur. Instead, shocks to one sector will generally generate a *substitution* of factors to one sector from the others. Hence, the predictions of the two models with regards to employment fluctuations are quite different.

Drawing on these implications, we investigated detrended data on employment fluctuations. Our primary finding was positive comovement in sectoral employment over the cycle which does not appear to be due simply to sectors experiencing common shocks. We interpret these findings as supportive of both the model of imperfect competition explored in Section I of the paper and a version of the Long-Plosser model in which technological complementarities dominate. These results do appear to be contrary to the predictions of many simple real business cycle models in which sector-specific shocks induce the sectoral reallocation of labor.

Our analysis leaves open a number of interesting and important questions. First, we have made some strong assumptions about the structure of demands: firms consume only the nonproduced good while workers consume both types of goods. An important extension of our model would therefore be to derive these differences explicitly. One avenue would be to allow the firm to be an infinitely lived institution. Shareholders could live for finite periods of time and be able to operate in capital markets while workers would live for a single period. In this setting, the nonproduced good could act as a store of value for shareholders, that is, it could simply be money so that the demand for this good was a means of obtaining future consumption. Owing to their single period of life, which might be viewed as a proxy for exclusion from capital markets, workers would have no demand for this good.[26]

Second, we have not really incorporated into our analysis any intertemporal strategic interactions. While these are apparently not necessary to generate the time-series and cross-sector variations of interest in this paper, these interactions by themselves might be worthy of attention.

Third, firms hold inventories in our model as a store of value. While this is analytically convenient, this is unlikely to be the only (key) reason for inventory holdings. In future work, we plan to integrate into our analysis a more reasonable rationale for inventory holdings along the lines of production bunching due to non-convexities in technology.[27] This setting is described by Valerie Ramey (1988) as capturing the technology of a number of industries and implies that production is more volatile than sales as suggested by recent empirical evidence (see, for example, Blinder, 1986). We conjecture that the dynamic effects of inventories will remain even in a setting with a richer explanation for inventory holdings.

Finally, a full comparison of this approach with that of Long-Plosser to understanding comovement across sectors seems warranted and would best be achieved by the construction of a model which incorporated in it markets for both final goods and intermediate products. This would allow us to evaluate the role of imperfect competition at each stage of the production process and to, perhaps, explore the casual linkages between sectors linked together by both factor and final goods demands.

APPENDIX: PROOF OF PROPOSITION 1

First, we must demonstrate that if $\epsilon_1 = 0$ and $M^i = M$, then (4) and (16) will characterize the prices and wages in an underemployment equilibrium. Because of the homothetic preferences, the price of each good in each period will be the same. Hence the wage rates must be the same each period for (16) to hold. Since marginal revenue is assumed to be a monotonically increasing function of price, prices will be the same each period.

---

[26] With this approach, firms shareholders are "life cycle consumers" while workers are "Keynesian consumers" because of credit rationing. In fact, this theme of credit rationing appears quite frequently in the literature on demand linkages across sectors since this provides a basis for increased production in one sector to increase spending in another. Empirically, this is supported by the work of Stephen Zeldes (1989) which provides empirical evidence that liquidity constraints are important for explaining consumption behavior. Mark Dynarski and Steven Sheffrin (1987) argue that the consumption behavior is dependent upon an agent's employment status which is consistent with the proposition that unemployed workers are liquidity constrained.

[27] One possibility we explore in Cooper-Haltiwanger (1989) is a production bunching model in which firms incur a fixed cost of "starting up the plant" and a low marginal cost up to capacity.

We must verify that (w^*, p^*) solving (4) and (16) exist. Figure 3 is a graphical representation of (4) and (16). The derivative of marginal revenue, $d[p(1+\xi(p))]/dp$, is assumed to be strictly positive so that the graph of (4) has a strictly positive slope for all p. As $p \to 0$, so must w. Otherwise, the worker would obtain infinite utility from positive w at $p=0$. The slope of (16) is the workers level of spending on the produced good which goes to infinity as $p \to 0$ and approaches 0 as $p \to \infty$. These two limit properties plus the quasi-convexity of the indirect utility function implies that there will be at least one crossing in Figure 3.

To show that the left side of (17) expresses output in each period for each sector, note that (13) is satisfied at $I_1 = 0$ so that the sectors are identical over time. Hence the solution will be symmetric and if (17) is satisfied, the quantities demanded can be produced so that an underemployment equilibrium will occur. Clearly, this condition is more likely to be satisfied if N is large and M small.

REFERENCES

Abraham, Katharine and Katz, Lawrence, "Cyclical Unemployment: Sectoral Shifts or Aggregate Disturbances?," *Journal of Political Economy*, June 1986, *94*, 507–22.

Arvan, Lanny, "Some Examples of Dynamic Cournout Duoploy with Inventory," *Rand Journal of Economics*, Winter 1985, *16*, 569–86.

Blinder, Alan, "Retail Inventory Behavior and Business Fluctuations" *Brookings Papers on Economic Activity*, 2:1981, 443–505.

______ "Can the Production Smoothing Model of Inventory Behavior be Saved?" *Quarterly Journal of Economics*, August 1986, *101*, 431–53.

______ **and Fischer, Stanley,** "Inventories, Rational Expectations and the Business Cycle," *Journal of Monetary Economics*, November 1981, *8*, 277–304.

Cooper, Russell, "Optimal Labor Contracts, Imperfect Competition and Underemployment Equilibria: A Framework for Analysis," NBER Working Paper No. 2060, 1986; *Canadian Journal of Economics*, forthcoming.

______, "Dynamic Behavior of Imperfectly Competitive Economies with Multiple Equilibria," NBER Working Paper No. 2388, 1987.

______ **and John, Andrew,** "Coordinating Coordination Failures in Keynesian Models," *Quarterly Journal of Economics*, August 1988, *103*, 441–63.

______ **and Haltiwanger, John,** "Macroeconomic Implications of Production Bunching: Factor Demand Linkages," NBER Working Paper No. 2976, January 1989.

Darby, Michael, Haltiwanger, John and Plant, Mark, "Unemployment Rate Dynamics and Persistent Unemployment Under Rational Expectations," *American Economic Review*, September 1985, *75*, 614–37.

Diamond, Peter, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, *90*, 881–94.

Dynarski, Mark and Sheffrin, Steven, "Consumption and Unemployment," *Quarterly Journal of Economics*, May 1987, *102*, 411–28.

Eichenbaum, Martin, "Rational Expectations and the Smoothing Properties of Inventories of Finished Goods,"*Journal of Monetary Economics*, July 1984, *14*, 71–96.

Hall, Robert, "Market Structure and Macroeconomic Fluctuations," *Brookings Papers on Economic Activity*, 2:1986, 285–322.

______, "The Volatility of Employment with Fixed Costs of Going to Work," Stanford University, mimeo., 1987.

______, "A Non-competitive, Equilibrium Model of Fluctuations," NBER Working Paper No. 2576, May 1988.

Haltiwanger, John and Maccini, Louis, "Inventories, Orders, Temporary and Permanent Layoffs: An Econometric Analysis," *Carnegie-Rochester Conference Series on Public Policy*, Spring 1989, 301–66.

Hansen, Gary, "Indivisible Labor and Business Cycles," *Journal of Monetary Economics*, November 1985, *16*, 309–27.

Hart, Oliver, "A Model of Imperfect Competition with Keynesian Features," *Quarterly Journal of Economics*, February 1982, *97*, 109–38.

Heller, Walter, "Coordination Failure Under Complete Markets with Applications to Effective Demand," in Equilibrium Analysis, Essays in Honor of Kenneth J. Arrow, Volume II, Walter Heller, Ross Starr, and David Starrett, eds., Cambridge, Cambridge University Press, 1986.

______, "Coordination Failure with Temporary Equilibria," Notes presented at 1987 NBER Summer Institute, July 1987.

Howitt, Peter, "Transactions Costs in the Theory of Unemployment," *American Economic Review*, March 1985, 75, 88–101.

Kydland, Finn and Edward Prescott, "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Lilien, David, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, *80*, 777–93.

Long, John and Plosser, Charles, "Real Business Cycles," *Journal of Political Economy*, February 1983, *91*, 39–69.

______, "Sectoral vs. Aggregate Shocks in the Business Cycle," *American Economic Review: Papers and Proceedings*, May 1987, *77*, 333–36.

Lucas, Robert, "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, *4*, 103–24.

______, "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, December 1975, *83*, 1113–44.

______, "Understanding Business Cycles," in *Stabilization of the Domestic and International economy*, K. Brunner and Allan H. Meltzer, eds., Carnegie-Rochester Conference Series on Public Policy, *5*, Amsterdam: North-Holland, 1977.

Ramey, Valerie, "Non-Convex Costs and the Behavior of Inventories," mimeo., University of California, San Diego, June 1988.

Rogerson, Richard, "An Equilibrium Model of Sectoral Reallocation," *Journal of Political Economy*, August 1987, *95*, 797–809.

Rotemberg, Julio and Saloner, Garth, "The Cyclical Behavior of Inventories," *Quarterly Journal of Economics*, February 1989, *104*, 73–98.

Sargent, Thomas, "Interpreting Economic Time Series," *Journal of Political Economy*, April 1981, *89*, 213–48.

Sims, Chris, "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

U.S. Department of Labor, Bureau of Labor Statistics, Employment and Earnings, (computer tape), 1987.

Weitzman, Martin, "The Simple Macroeconomics of Profit Sharing," *American Economic Review*, December 1985, *75*, 937–53.

Zarnowitz, Victor, "Recent Work on Business Cycles in Historical Perspective," *Journal of Economic Literature*, June 1985, *23*, 523–80.

Zeldes, Stephen, "Consumption and Liquidity Constraints: An Empirical Investigation," *Journal of Political Economy*, April 1989, *97*, 305–46.

Did the Federal Trade Commission's Advertising Substantiation Program Promote More Credible Advertising?

By RAYMOND D. SAUER AND KEITH B. LEFFLER*

This paper examines the effects of the Federal Trade Commission's Advertising Substantiation Program, developed in the early 1970s. This program coupled changes in the legal definition of deception with more vigorous FTC enforcement. We analyze changes in advertising intensity, media choice, media wealth, and the progress of new entrants. The evidence suggests that adoption of substantiation requirements increased the credibility of advertising. (JEL 619,531)

Advertisers have an obvious incentive to misrepresent their products, an act which is restrained by market forces. These forces give rise to observed institutions such as outside verification and nonproducer supplied information, in addition to incentives which directly motivate honesty on the part of the producer himself.[1] Both the institutions and the misrepresentations themselves are costly. Intervention by government will change outcomes in such a world, but it is not clear what outcomes should be expected. George Stigler (1971) reminds us that the effects of a policy may not coincide with the stated intent of its proponents; more sanguine conclusions emerge from Gary Becker's (1983) model of political competition.

Intervention in advertising increased markedly in the early 1970s when the Federal Trade Commission implemented its Advertising Substantiation Program (ASP). In this paper we examine whether the FTC achieved its stated goal of motivating more truthful advertising. We discuss the regulatory background in Section I. In Section II we present a variety of evidence on changes in advertising intensity, media mix, media wealth, and the process of new product entry. In every case we find the data to be consistent with an increase in the credibility of advertising claims. Section III concludes the paper.

I. Regulatory Background

By all accounts, the Federal Trade Commission was a woefully ineffective monitor of advertising prior to 1970. To some observers, this meant that the FTC's hand needed to be strengthened by devoting more resources, changing the rules of the game, or both. *The Nader Report* (1969) and the *ABA Report* (1969) put forward this view. To others—notably Yale Brozen (1973), Richard Posner (1969, 1973), and Stigler[2]—ineffectiveness (at best) was to be expected, and additional resources in the hands of the FTC were likely to be counterproductive.

Consumerist pressure succeeded in revamping the FTC. The agency was reorganized, 18 of 31 top-level staff members were discharged, and 70 percent of the existing advertising cases were dropped in preparation for a new regulatory initiative.[3] A land-

---

*Department of Economics, Clemson University, Clemson, SC 29631 and Department of Economics, University of Washington, Seattle, WA 98195. We are grateful to Yoram Barzel, Bill Dougan, Jim Ferguson, Mason Gerety, Levis Kochin, Matt Lindsay, Ed Rice, Sam Peltzman, Tim Sass, and seminar participants at Clemson, New Mexico, and Texas for their many helpful comments and suggestions. We also thank the referees, whose very useful comments helped improve the paper.

[1] Michael Darby and Edi Karni (1973), Benjamin Klein and Keith Leffler (1981), and Carl Shapiro (1983) analyze these issues.

[2] A Stigler talk on consumerist motivated regulation is covered in *Advertising Age*, 3/8/71, p. 14.

[3] See Kenneth Clarkson and Timothy Muris (1981), and *Advertising Age*, especially 12/8/69, p. 1, for ac-

mark complaint was issued in April 1970 against Pfizer Corporation. This case became a successful test of a new, stringent requirement of the FTC upon advertising: that claims of a verifiable nature must have *prior*, fully documented, and adequate substantiation.[4] Acceptance of the substantiation doctrine shifted the burden of proof from the FTC to advertisers, and thereby reduced the costs to the FTC of enforcement.

Monitoring activity increased sharply. Demand letters for substantiation were sent to targeted advertisers, a policy which had the dual role of generating new cases and informing the industry of the new regime. New cases increased from 1 each in 1968 and 1969 to 6 in 1970, 13 in 1971, and 24 in 1972. This flurry of regulatory activity was a hot issue in the advertising trade press. But did it change the behavior of advertisers?

II. Empirical Analysis of the Advertising Substantiation Program

For the FTC's efforts to impact favorably information in advertising, two criteria must be satisfied. The agency must (i) have the power to deter and (ii) it must use this power to establish appropriate incentives for advertisers. Sam Peltzman (1981) provided evidence that the FTC has a powerful club indeed. In a five-day period surrounding the announcement of a false advertising case, his sample of firms lost 3 percent of their market value. We found losses of similar magnitude for the population of all substantiation cases issued against publicly traded firms.[5]

The first criterion satisfied, we now explore the possibility that ASP succeeded in motivating more truthful ads. We do not propose to somehow measure truth. Instead we proceed by asking what would be observed if the FTC were successful? Our hypothesis is that consumer scepticism of ad claims in general is positively related to the frequency of false claims. If the FTC is able to reduce this frequency, scepticism will decrease and therefore the return to making verifiable claims will rise. We call this effect an increase in the credibility of ad claims. We test for its presence in the remainder of the paper. We begin by applying this logic to advertising intensity. Subsequent applications involve media choice and new product entry.

A. *Advertising Intensity*

Our analysis emphasizes the differential impact of substantiation requirements across different types of goods. We define type I goods as goods for which information on various attributes is both valued and costly for the consumer to obtain. Infrequently purchased durables would be prototypical of such goods. Type II goods are those for which the relevant information is either known or can be cheaply obtained via purchase and consumption. Most inexpensive and frequently purchased goods are of this type.

The credibility of producer claims regarding type II goods should be effectively monitored by market forces. Not only is information about the good cheaply gained through consumption; high purchase frequency imposes a penalty on unsatisfactory performance by the seller. In contrast, type I goods are precisely those for which consumers value more credible information, and for which infrequency of purchase dilutes the incentive to motivate honest representations.

---

counts of this transformation of the FTC. Dorothy Cohen (1980), and Collot Guerard and Julie Niemasik (1978), discuss the development of ASP in particular.

[4] The complaint charged that Pfizer's claim that the product Un-burn "actually anesthetizes nerves in sensitive sunburned skin" was deceptive because Pfizer lacked adequate substantiation for the claim. This particular claim may be verifiable; substantiation is not required for inherently non-verifiable claims ("there's something about an Aqua Velva Man") or for puffery ("we're number one for everyone"). We discuss Pfizer in Section II; see especially fns. 11 and 12.

[5] Using event study techniques, we calculate an average loss of 3 percent ($t = 4.3$) for 31 publicly traded firms issued substantiation complaints. For the 24 cases not in Peltzman's sample, the average loss was 1.7 percent ($t = 2.9$). These costs can be enormous: the average loss for the *smaller* sample exceeds $10 million in current (1990) dollars.

If ad claims are more credible, they are of greater value. We expect advertising to increase with increases in its value. Substantiation is designed to improve the credibility of claims, but this comes only at a cost. Hence, we argue that the direct effects of ASP, both costs and benefits, apply mainly to type I goods, for which informative advertising is more highly valued. Producers of type II goods can simply avoid the costs of substantiation by not making verifiable claims, that is by "puffing," with little or no adverse sales effect. Thus, if credibility effects are large relative to substantiation costs (an effective ASP), we expect advertising of type I goods to rise relative to type II goods. On the other hand, if credibility effects are absent (or small relative to substantiation costs), advertising of type I goods should fall.

To test this hypothesis we examine firm-level advertising intensities for infrequently purchased durables (type I goods) and for inexpensive, frequently purchased products (type II goods). Leading National Advertisers (LNA) provides data on firm-level advertising expenditures in national media. We chose firms from eight LNA industry classes which could be unambiguously categorized as type I or type II for study.[6] The four type I (infrequently purchased durable goods) industries are automobiles, furniture, household equipment, and office equipment. The four type II industries (inexpensive, frequently purchased items) are cosmetics and toiletries, foods, soft drinks, and sweets. Sales data from Standard & Poor's *Industry Studies* were combined with advertising expenditures from LNA to calculate advertising intensities (ad expenditures divided by sales) for the 18 type I firms and 20 type II firms listed in both sources.

Pre-ASP intensities were calculated using data for 1967 and 1968. We cannot use data for 1969 since LNA's data on television expenditures is incomplete for this year. Choosing the post-substantiation period is problematic. Presumably it takes some time for credibility changes to take effect; however, advertising is influenced by many factors which may change over time, which augurs against choosing a period many years after the base period. The regulatory efforts in 1970 and 1971 were widely reported in the trade press and events in the first six months of 1971 (described in the next section) clearly defined the new regime. Hence, we choose 1972 and 1973 as post-substantiation years.

The average change in ad intensity is listed by industry in the first column of Table 1. We find a uniform decline in ad intensity for the type II industries, with very little change for type I.[7] Indeed, the largest decline among the type I industries (furniture) is but a third of the *smallest* decline among the type II industries (foods). These changes are sufficient to raise the average ad intensity of the type I firms by 30 percent relative to the type II firms, a rather substantial change.

The statistical significance of these changes is confirmed with a regression on the panel data set of firms. In Table 2 we present least squares estimates of the equation

$$ADIN_{it} = \alpha_i + \beta_1 \cdot \mathrm{ASP}_t + \beta_2 \cdot \mathrm{ASP}_t \cdot TYPEI_i,$$

where $ADIN_{it}$ denotes the ad intensity of firm i in year t, α_i is an intercept specific to firm i, ASP is a 0–1 variable taking the value of 1 for years 1972 and 1973, and *TYPEI* similarly takes the value of 1 for type I firms. An increase in ad credibility is consistent with a positive value for the coefficient β_2. The estimate of β_1 will capture the effects of all omitted factors which changed over the period, in addition to picking up general effects of ASP.

The differential effect for type I firms (the coefficient β_2 in Table 2) is seen to be positive and statistically significant ($t = 4.68$).

---

[6] Such an analysis depends critically on the classification scheme. We chose to focus on those industries that could be unambiguously classified rather than make an attempt to cover all industries. The alcohol, pharmaceutical, gasoline, and tobacco industries were not studied due to extensive regulation unconnected with ASP. Classification ambiguities ruled out apparel, publishing, travel, and insurance.

[7] Appendix A lists the changes in intensity across the period for each firm.

TABLE 1—ADVERTISING INTENSITY AND MEDIA CHOICE

Type I Goods:	Change in AD Intensity	Change in TV Percentage
Appliances	−0.0008	0.1234
Automobiles	−0.0007	0.0768
Furniture	−0.0029	0.1183
Office Equipment	0.0002	0.1888
Mean of 18 Firms	−0.0010	0.1243
Type II Goods:		
Cosmetics	−0.0136	0.0675
Foods	−0.0080	−0.0416
Soft Drinks	−0.0305	0.0005
Sweets	0.0159	−0.0304
Mean of 20 Firms	−0.0177	0.0005

Notes to Table 1: The firms used for each industry are listed in Appendix A, available from the authors on request. Ad intensities for each firm are the ratio of national advertising expenditures listed for each firm in LNA's *National Advertising Investments* to net sales listed in Standard and Poor's *Industry Studies*. The firms chosen were all firms listed under the industry classification in *Industry Studies*. TV intensity is the percentage of total advertising during each period accounted for by network and spot television expenditures. Changes are calculated for the period 1972–73 relative to 1967–68. LNA data are incomplete for 1969 which precluded the use of that year.

TABLE 2—PANEL DATA REGRESSIONS OF AD INTENSITY AND TELEVISION PERCENTAGE

Dependent Variable:	*ADINT*	*TVPERC*
α_i	√	√
ASP	−0.0187	0.0005
	(7.18)	(0.02)
TYPEI·ASP	0.0177	0.1238
	(4.68)	(3.53)
R^2	0.32	0.18
OBSERVATIONS	152	152

Notes to Table 2: The numbers in parentheses are t-statistics. Both equations include 38 firm-specific intercepts (α_i). The R^2 reported above are 1-*SSE/SST*, where *SST* is the sum of squared residuals from a regression of the dependent variable on the firm-specific dummies alone. The R^2 statistics from the regressions above including the effects explained by the firm dummies are 0.93 and 0.87, respectively. There are 4 observations (1967, 1968, 1972, 1973) for each of 18 type I and 20 type II firms, yielding 152 observations for each regression.

More general empirical studies (Issac Erlich and Lawrence Fisher, 1982; Ward Bowman, 1976, and Timothy Bresnahan, 1986) document that media prices, scale economies, and consumers' opportunity costs influence ad intensity. Real sales and media prices both increased during this period, which can account for the negative estimate of β_1 (i.e., some of the observed decline in ad intensity of type II goods).[8] Still, it is the differential change that is relevant to our theory; both the panel data and the industry averages identify a relative increase in the ad intensity of type I goods coincident with ASP. This evidence is thus consistent with the hypothesis of an increase in the credibility of advertising.

If these changes reflect increases in the value of advertising, one would expect firms relying on advertising to enjoy increased wealth. Richard Higgins and Fred McChesney (1986) examined wealth effects on large advertisers during this period. They concluded that "substantiation had a positive effect on the value of the market leaders," which corroborates the interpretation that advertising value was enhanced by substantiation.

---

[8] We also estimated regressions including media prices and the manufacturing wage (to proxy opportunity costs as in Erlich and Fisher, 1982), both in real terms. The coefficient on the differential effect of ASP in the presence of the real price of advertising remains highly significant; the coefficient on ASP becomes positive and insignificant. Erlich and Fisher (1982) estimate separate derived demand equations for advertising of durable and nondurable goods, a fairly close parallel to our type I and type II classification. The coefficient estimates for sales are virtually identical (see Erlich and Fisher, p. 383) and exhibit scale economies; since sales growth for the two industry types are also virtually identical, scale economies in advertising can account for an overall decline in intensity, but not the differential increase of type I goods. On the other hand, the estimated price response of type I goods is greater than that of type II; hence the Table 2 specification is biased *against* finding a credibility effect of ASP since the price of advertising increased during the period.

B. *Media Effects*

Substantiation will also have differential effects across media. "Facts" may be easily stated on television, but such statements will be of little value in a market with misleading claims and the consequent consumer skepticism. In print however, such skepticism can be overcome by voluntarily providing substantiation. This is far more costly to do on television, in which viewers of the ad are not self-selected, and in which time to present and absorb details is more scarce.[9] Consequently, if FTC action were to increase the credibility of factual claims, this would have its greatest effect in the television medium, where unadorned claims would benefit most from outside verification.

Furthermore, the effects should be greater for type I goods, where credibility is of greater value. Therefore an additional hypothesis is that under an effective substantiation requirement, the use of television advertising should rise for type I goods relative to type II goods. We calculated the percentage of total advertising expenditures spent in television (*TVPERC*) for each firm using data from LNA. The second column of Table 1 contains the average change in *TVPERC* for each industry. We see that the use of television increases uniformly among the type I goods, with a mix of signs for type II. The second column of Table 2 presents estimates from the panel data regression of *TVPERC* on ASP and ASP$\cdot$*TYPEI* at the firm level. Again, we see a significant differential increase for type I firms ($t = 3.53$). This is what one would expect if substantiation requirements helped to overcome a credibility problem in the television medium.

If credibility effects occurred, the increased demand for advertising space on television should have led to an increase in the price of TV space relative to other media. This is indeed the case. The McCann-Erickson ad agency constructs media price indexes for television, newspapers, and magazines. The indexes are, for 1970: 113, 115, 109; and for 1973: 140, 128, 110, respectively. Hence the price of television space rose roughly 24 percent in this period, a good bit more than newspapers (11 percent) and magazines (1 percent).

We have further examined this issue with an event study which tests for substantiation-generated wealth effects on media firms. If market participants thought that increased monitoring of advertising would help the television medium overcome a credibility problem, one would observe a positive differential abnormal return between stock portfolios of the television and print industries.[10] We analyze four important substantiation events: April 14, 1970, the date the first substantiation case (Pfizer) became publicly known; Jan. 21, 1971, when the FTC announced that Pfizer was to be a test case for the substantiation doctrine; April 16, 1971, the date the hearing examiner in Pfizer issued a ruling affirming the substantiation rule; and June 11, 1971, when FTC issued in its "boldest move yet" the substantiation regulation.[11]

---

[9] This may be why television ads occasionally refer viewers to "your favorite weekly magazine" for more detailed information on the product.

[10] Mean cumulative abnormal returns (*MCARs*) for the two portfolios were calculated with the market model, described in Appendix B. Media stocks were identified from a 1971 advertising stock price listing in *Ad Age*. SEC 10-K reports were analyzed to determine if a company's revenues were derived primarily from television broadcasting or print. ABC, CBS, RCA, (NBC), Cox, Metromedia, Storer, and Wometco were classified in the TV group. Gannett, *Esquire*, *Time*, Meredith, Knight, *New York Times*, and *Times-Mirror* were classified in the print group.

[11] See "Test Case Opens on Un-Burn Ads" in *Advertising Age*, 1/25/71. The hearing examiner's decision, while exonerating Pfizer, did so because the claim was

TABLE 3—MEAN CUMULATIVE ABNORMAL RETURNS TO MEDIA STOCKS (PERCENT CHANGES IN THE PORTFOLIOS)

	Periods						
	(−40, −1)	(−20, −1)	(−5, −1)	(0,4)	(0.19)	(0,39)	(−3,1)
I. The Pfizer Case Announcement: 4/14/70							
MCARTV	1.82	−1.41	0.07	−2.01	-8.67^{b}	−7.24	−3.01
MCARP	−3.37	−0.46	−0.52	0.86	−5.95	-12.62^{a}	0.50
DMCAR	5.19*	−0.95	1.22	−2.87	−2.72	5.38*	−3.51
II. The Test Case Announcement: 1/21/71							
MCARTV	5.30	5.04	2.24	−0.39	−0.71	10.01	2.08
MCARP	1.42	−2.12	−3.57	−1.04	−0.89	4.17	−3.02
DMCAR	3.88	7.16***	5.81***	0.65	0.18	5.84*	5.11***
III. The Decision Affirming Substantiation: 4/16/71							
MCARTV	12.91^{a}	4.02	1.31	1.59	3.53	8.91	0.15
MCARP	2.87	−1.42	−1.15	2.74	3.40	3.01	0.94
DMCAR	10.04***	5.44**	2.46**	−1.16	0.13	5.90**	−0.79
IV. Announcement of the Substantiation Regulations: 6/11/71							
MCARTV	12.34^{b}	8.07^{b}	0.44	−0.08	−1.82	1.00	−1.40
MCARP	3.63	−1.54	0.57	−0.48	−1.23	0.10	0.89
DMCAR	8.71***	9.61***	−0.13	−0.40	−0.59	0.90	−2.30

***: significant in a one-tailed *t*-test at the 0.01 level.
**: significant in a one-tailed *t*-test at the 0.05 level.
*: significant in a one-tailed *t*-test at the 0.10 level.
^{a}: significant in a two-tailed *t*-test at the 0.05 level.
^{b}: significant in a two-tailed *t*-test at the 0.10 level.

Notes to Table 3: Our hypothesis refers to differences in the abnormal returns between the two portfolios. Hence we apply one-tailed tests for the difference and two-tailed tests to the individual portfolios. *MCARTV* is the mean cumulative return to the television portfolio, *MCARP* to the print portfolio, and *DMCAR* is the difference between the two. All estimates are based on the market model of Fama (1976), using data from the *CRSP* tapes. The construction of these returns and their standard errors are described in Appendix B.

Table 3 reports summary statistics on the differential abnormal returns for these events. It appears that Pfizer was not anticipated, as there is no sign of abnormal returns prior to the announcement.[12] Negative returns to both portfolios follow the announcement of the Pfizer case, suggesting a market belief in the cost increasing aspect of substantiation.

A consistent pattern of preannouncement positive differential returns to the TV portfolio occurs over the next three events. In the week before the next two events, we observe significant positive differential returns to the TV stocks. In all three events, the one-month abnormal return prior to the announcement is significantly greater for the TV stocks; the cumulative differential return is a hefty 22.2 percent for these periods (two of these

---

found to be substantiated. His decision took special care to affirm the legality of the FTC's substantiation requirement. See 81 *FTC Decisions* 53, and *Advertising Age*, 5/10/71. The regulation issued in June required firms to produce substantiation on demand of the FTC. See "FTC to Require Proof of Public Ad Claims: Unprecedented Action Falls Just Short of Nader Idea" in the *Wall St. Journal*, 6/11/71.

[12] This is confirmed by the response of Pfizer stock. The abnormal returns for Pfizer are −0.1 percent for week prior to the case announcement, and −4.6 percent for the week beginning on the announcement date. Hence, the decline does not occur until the event date, suggesting that Pfizer was a surprise. This differs from other cases in both our analysis and that of Peltzman (1981) in which we observe losses in the days prior to the announcement.

months are significant at the 0.01 level, the third at the 0.05 level).

We have looked for other potential causes of these return differentials by canvasing the *Wall Street Journal Index* and the issues of *Advertising Age* during this period. Several articles document the FTC's reforms and industry responses, while no rival explanation for these changes is apparent.[13] These returns thus complement the evidence reported above, which suggests that substantiation of advertising claims helped television overcome a credibility problem.

C. *Did Informative Advertising Really Increase?*

If the changes reported above were due to increased credibility of advertising claims, and not some other source, the mix of advertising would shift to include more informative claims. Hence it would be useful to know if this took place. There are obvious problems with classifying ad messages as informative, noninformative, and the like.[14] Nevertheless, studies in the marketing literature do exactly that. John Healey and Harold Kassarjian (1983) had student volunteers classify informational signals in magazine advertisements in an attempt to evaluate ASP. It is difficult to draw strong conclusions from their study, but on one point the data speak loudly. In a three-part partitioning of claim verifiability into verifiable, perhaps verifiable, and non-verifiable, claims shifted from the center category toward the two ends of the spectrum in the post-substantiation period. The classifiers found that the post-ASP proportion of verifiable claims was 17 percentage points higher; non-verifiable claims increased 10 points. This suggests to us that substantiation entailed both significant benefits and significant costs: vague claims were less likely to be made in favor of either fully substantiated claims or claims that were non-verifiable and hence not subject to substantiation.

D. *Product Entry and Substantiation Requirements*

Advertising is of particular value for new products. An early charge leveled against substantiation requirements was that they would increase the costs of entry and hence limit competition (see Brozen, 1973). This reflects the a priori belief that credibility effects would be negligible and substantiation costs large. If the costs dominate, fewer verifiable claims would be made, and entrants would encounter greater difficulty matching their product with consumers. It would thus take longer for entrants to reach their ultimate sales potential. Alternatively, if credibility effects dominate, consumers would respond with greater force to verifiable claims. Matching consumers to products would be less costly, repeat purchase probabilities would increase, and entrants would approach their sales potential more rapidly.[15]

We examine the issue of substantiation and entry with monthly sales data for all new products entering the over-the-counter drug market in the years 1967–69 and 1973–75.[16] Our aim is to characterize the rate of progress of the post-ASP entrants

---

[13] Note that general trends which may increase the value of TV ads would *not* generate the portfolio effects we observe in an efficient markets world. The only notable story in or near a significant event window was a report on proposed legislation requiring substantiation, in the 4/5/71 issue of *Advertising Age*.

[14] David Laband (1989) discusses these problems. It is worth noting that his classification scheme indicates a much greater incidence of informative signals in the television medium than was reported by previous researchers. In addition, he presents an economic model which successfully explains variation of informational signals across media, which lends some support to his classification scheme.

[15] More credible claims would increase trial purchase probabilities. However, the purpose of false advertising is presumably to attract trial purchases of a product. Hence, the credibility effect on trial purchase probabilities is ambiguous, since it depends on decreasing the frequency of false claims.

[16] The data are, specifically, the monthly value of shipments to drugstores of new products in the OTC pharmaceutical market as reported in *InterContinental Medical Statistics*. The data vary in coverage from 16 to 24 months from initial entry.

TABLE 4—ENTRY PATHS PRE- AND POST ASP PERCENT OF MATURE SALES LEVEL OBTAINED BY MONTH

	4	8	12	16	AVG. Growth Coefficient
Average Pre-ASP 67–69	65.6	70.5	64.6	66.5	0.986
Average Post-ASP 73–75	52.1	59.3	81.2	77.9	1.387
$T(D)$					1.72**

**Significant in a one-tailed test at the 0.05 level.

Notes to Table 4: $T(D)$ is the t-statistic which tests the null hypothesis that the difference is random between the two periods. The percentage reported is the average for each sample of actual sales at month t for entrant i of the estimated mature sales level for entrant i. Estimates of the mature sales level and the growth coefficient for each product were obtained by OLS estimation of equation (7) in the text. Summary statistics, including the firm-level components of this table are reported in Appendix C, available from the authors on request.

relative to their pre-ASP counterparts. To do this requires estimates of each product's sales potential in addition to simple growth rates of their sales. Both of these can be obtained by estimating the parameters of a logistic growth curve. This function implies that the sales path is s-shaped which seems quite appropriate for new product entry. The logistic function can be transformed into the following linear form:

$$PCT\Delta Y_{it} = g_i - (g_i/K_i) * Y_{it}$$

where Y_{it} and percent ΔY_{it} are the level and percentage change in monthly sales of entrant i, g_i is an estimate of the growth coefficient of the logistic curve, and K_i is the "mature sales level" of the product, which is easily calculated as the ratio of the first to the second coefficient estimate.

Results from estimating the logistic function for each product are presented in summary form in Table 4.[17] The average growth coefficient is significantly higher in the post substantiation period (the t-statistic for the difference in the means is 1.72, which is significant at the 0.05 level). In addition, we calculate the average (actual) sales position of entrants relative to the estimated mature sales level at different points on the entry time line as:

$$PCT_t = AVG(Y_{it}/\hat{K}_i).$$

These percentages, calculated for the pre- and post-ASP entrants at 4, 8, 12, and 16 months after entry, are also presented in Table 4. The entry path for the post-ASP entrants is steeper, with a steadier approach toward the mature sales level, indicating higher repeat sales. These data thus exhibit no deleterious effect of substantiation requirements on new product entry. Rather, the evidence is once again consistent with ASP increasing the credibility of advertising.

III. Conclusion

In their in-house evaluation of ASP, Guerard and Niemasik (1978, p. 4) note that "The advertising community appears to have modified its behavior. We no longer see patently false or deceptive claims as was the case a number of years ago...and it is rare when we find a company that lacks any substantiation for its claims." Economists quite properly take such statements with a grain of salt. Being skeptics at the outset, we were surprised to find each test in this paper consistent with an increase in the credibility of advertising. It is clear however that the substantiation doctrine, by shifting the burden of proof to advertisers, decreased the FTC's cost of enforcement. Peltzman showed, and we confirmed, that heavy penalties from the capital market await firms caught in the FTC's net. With these facts in mind, under any interpretation of the FTC's motives, changes of some form will occur due to increases in FTC monitoring activity. The changes we observe in a number of settings consistently point to the conclusion that the FTC succeeded in promoting more credible advertising during this period.

---

[17] Appendix C, available from the authors on request, lists coefficient estimates and summary statistics from the 103 individual product regressions. The coefficient estimates are significant at the 0.05 level for almost all products. The range of explanatory power is quite large however, with R^2 ranging from 0.00 to 0.94. Thus, some entrants do not fit the logistic mold. We do not exclude the nonconformists in order to keep the sample free of selectivity bias. Entrants for which data exist which were not used were suntan lotions (due to seasonal sales), and those entrants who failed to achieve monthly sales of $5000 for a minimum of three months.

APPENDIX A

Firms Corresponding to Categories in Table 1

Industry/Company	Δ*Ad Int*	ΔTV Intensity
Appliances:		
Maytag	0.0026	0.3321
Singer	−0.0006	0.0867
Sunbeam	−0.0062	0.0187
Whirlpool	−0.0027	0.1295
Zenith	0.0027	−0.0501
Autos:		
AMC	−0.0001	−0.0560
Chrysler	−0.0014	0.0943
Ford	0.0001	0.1262
GM	−0.0014	0.1425
Furniture:		
Bassett	−0.0026	0.3874
Henredon	−0.0028	0.0000
Kroehler	−0.0025	0.2162
Simmons	−0.0038	−0.1305
Office Equipment		
Control Data	0.0033	0.4131
Honeywell	−0.0009	0.0513
IBM	−0.0000	0.2704
NCR	−0.0008	0.0000
Sperry-Rand	−0.0006	0.2054
Cosmetics:		
Alberto	−0.0014	0.0276
Avon	−0.0072	0.1385
Chese-Ponds	−0.0184	0.0085
Gillette	−0.0360	−0.0436
Revlon	−0.0050	0.2063
Foods:		
Campbell	−0.0104	−0.1439
Castle	−0.0017	−0.3738
Del Monte	−0.0042	0.3025
Green Giant	−0.0181	−0.0497
Heinz	−0.0054	−0.1387
Soft Drinks:		
Coca-Cola	−0.0173	0.0221
Crush	0.0070	0.0018
Dr. Pepper	−0.0193	0.0944

APPENDIX A: CONTINUED

Industry/Company	Δ*Ad Int*	ΔTV Intensity
Pepsi	−0.0102	−0.0151
Royal Crown	−0.0566	−0.0926
Seven-Up	−0.0864	−0.0077
Sweets:		
Hershey	0.0186	−0.0473
Peter Paul	−0.0211	−0.0610
Tootsie	−0.0131	0.0134
Wrigley	−0.0481	−0.0266

Notes: The firms selected for study are all firms who (a) were advertisers in both the 67–68 and 72–73 period, and (b) had balance sheet reports printed in Standard and Poor's *Industry Studies*.

APPENDIX B

Event Study Methodology

We calculate abnormal returns using the market model of Fama (1976). The model is:

$$(\text{A1}) \qquad R_{it} = \alpha_i + \beta_i \cdot R_{mt} + \varepsilon_{it} \quad i = 1, 2, \ldots n,$$

where R_{it} is the return for firm i at t, R_{mt} is the return on the equally weighted portfolio of NYSE common stocks, obtained from the *CRSP* tape, and $t = 0$ is the announcement date.

Equation (A1) is estimated for each of the n firms in the sample over a one-year period (240 trading days), terminating 120 days prior to the announcement date.

These equations are used to generate prediction errors over the subsequent 80 trading days:

$$(\text{A2}) \qquad e_{it} = R_{it} - \hat{R}_{it},$$

where $\hat{R}_{it} = \hat{\alpha}_i + \hat{\beta}_i \cdot R_{mt}$,

is obtained using the OLS estimates of (A1). The variance of the prediction error is estimated by

$$\sigma_{ei}^2 = (1/80) \cdot \sum_{t=-120}^{-41} e_{it}^2.$$

The mean cumulative abnormal return (*MCAR*) over the period (t_b, t_e) for a sample of n firms is:

$$(\text{A3}) \qquad MCAR(t_b, t_e) = 1/n \cdot \sum_{i=1}^{n} \sum_{s=t_b(i)}^{t_e(i)} e_{is}.$$

The standard errors are:

$$(\text{A4}) \qquad s(MCAR[t_b, t_{se}]) = \left[1/n^2 \cdot (abs[t_b - t_e] + 1) \cdot \sum_{i=1}^{n} \sigma_{ei}^2 \right]^{1/2}.$$

The t-statistics used for testing the differences in the *MCAR*s for TV and print are calculated using the standard procedure for differences in means with different, unknown variance.

APPENDIX C

Entry Estimation Results for OTC Drugs Pre-ASP

Company	Growth	*Tg*	MSL	*Tm*	*RSQ*	*PCT*4	*PCT*8	*PCT*12	*PCT*16
1	2.201	2.01	8.484	1.49	0.24	0.471	0.825	0.000	0.000
2	0.974	3.61	6.929	3.10	0.25	0.722	0.289	0.722	0.722
3	1.227	2.28	144.517	2.01	0.12	0.920	1.944	0.740	0.623
4	0.340	2.10	23.334	2.19	0.13	0.514	2.057	0.857	0.643
5	0.330	2.55	24.038	2.16	0.14	0.666	0.458	0.541	0.790
6	1.146	3.29	11.394	3.18	0.27	1.229	1.053	0.439	0.263
7	1.512	3.74	17.044	3.32	0.29	2.112	1.291	0.352	0.704
8	1.959	2.05	704.364	1.71	0.10	1.440	0.237	1.393	0.463
9	0.315	3.26	141.215	3.77	0.36	0.453	0.779	0.885	0.744
10	1.173	3.25	151.285	3.13	0.29	0.370	0.932	0.873	0.826

APPENDIX C: CONTINUED

Company	Growth	*Tg*	MSL	*Tm*	*RSQ*	*PCT*4	*PCT*8	*PCT*12	*PCT*16
11	0.121	0.49	200.636	0.37	0.00	1.500	0.364	1.256	0.558
12	2.505	3.93	86.437	3.67	0.30	0.359	0.567	0.775	0.879
13	0.536	1.64	33.641	1.15	0.05	0.832	0.238	0.119	0.268
14	0.226	2.10	33.010	2.75	0.24	1.666	1.272	0.636	0.666
15	2.265	3.95	5.326	3.20	0.31	0.188	0.000	1.314	1.126
16	0.859	3.47	3.370	3.03	0.25	0.297	0.594	0.594	0.297
17	0.685	1.98	81.222	1.52	0.09	0.345	0.591	0.172	0.320
18	1.951	1.41	108.009	0.56	0.01	0.324	0.065	0.120	0.083
19	0.754	2.46	66.589	2.17	0.15	0.646	0.556	0.285	0.661
20	1.444	1.49	3.861	0.98	0.06	0.000	0.259	0.259	0.259
21	0.163	2.33	33.284	3.44	0.31	1.142	1.322	0.931	1.022
22	0.280	1.75	51.599	1.83	0.11	1.260	0.640	0.659	0.988
23	0.372	2.94	155.337	1.91	0.10	0.019	0.045	0.283	0.354
24	0.318	2.19	123.525	1.64	0.08	0.356	0.405	0.259	0.453
25	0.372	4.56	43.502	3.53	0.41	0.391	0.621	0.644	0.828
26	1.429	3.73	3.865	3.23	0.35	0.776	0.776	0.776	1.035
27	0.985	5.63	261.979	5.44	0.54	0.931	0.870	0.935	0.844
28	1.660	2.22	10.260	1.69	0.13	0.097	0.097	1.170	0.195
29	0.885	2.25	36.103	1.99	0.11	0.360	0.277	0.138	0.997
30	0.330	2.16	236.884	1.74	0.10	0.333	0.823	0.422	0.756
31	0.977	4.17	8.433	3.93	0.40	0.830	0.830	1.304	0.593
32	1.199	5.68	212.902	5.11	0.46	0.268	0.930	0.986	0.766
33	0.422	2.54	6.671	2.47	0.24	0.300	0.450	0.600	1.349
34	1.052	2.42	112.466	1.76	0.11	0.000	0.365	0.320	0.000
35	1.530	2.23	77.549	1.81	0.09	0.374	1.135	0.219	0.877
Average	0.986					0.671	0.705	0.6436	0.665

(See notes at the end of the appendix.)

			POST-ASP						
Company	Growth	Tg	MSL	*Tm*	*RSQ*	*PCT*4	*PCT*8	*PCT*12	*PCT*16
1	0.724	2.36	45.150	1.68	0.09	0.000	0.221	0.421	0.244
2	0.778	2.92	82.125	2.39	0.18	0.353	1.656	1.145	0.475
3	0.573	1.87	37.246	1.51	0.09	2.336	0.564	0.832	0.886
4	7.718	3.66	44.848	3.45	0.31	1.093	1.539	1.249	0.491
5	2.780	3.20	82.448	2.77	0.22	0.000	0.679	0.825	0.534
6	3.527	7.83	481.278	7.60	0.67	0.848	1.170	1.120	0.862
7	6.649	6.69	197.434	6.40	0.59	0.704	1.124	0.841	1.054
8	0.404	3.30	86.247	1.92	0.12	0.023	0.104	0.070	0.151
9	0.121	1.01	24.424	1.03	0.04	0.696	0.287	0.409	0.246
10	1.062	3.41	251.939	2.95	0.25	0.560	0.929	0.552	1.242
11	0.626	2.41	69.079	2.11	0.15	1.042	0.840	0.275	1.419
12	0.483	1.99	24.960	1.90	0.11	1.643	1.042	1.803	1.002
13	0.543	2.58	65.830	2.50	0.19	0.987	1.033	0.562	0.592
14	0.250	1.88	17.737	1.55	0.08	0.000	0.338	0.733	0.620
15	0.838	3.54	72.099	3.12	0.28	0.236	1.012	0.943	0.374
16	0.131	0.86	1.734	2.40	0.45	1.154	1.154	0.577	1.154
17	0.286	2.01	32.014	1.45	0.08	0.156	0.219	0.437	0.312
18	0.925	3.80	37.266	3.67	0.32	1.047	0.268	0.778	0.725
19	0.516	3.09	44.956	2.35	0.16	0.133	0.267	0.534	0.868
20	0.408	3.00	39.386	2.20	0.13	0.102	0.381	0.381	0.508
21	1.553	3.19	18.766	2.99	0.28	0.160	0.639	0.373	0.266
22	0.928	2.41	7.248	2.44	0.30	1.104	1.380	0.690	0.552
23	0.624	0.30	26.778	0.34	0.10	0.075	0.000	0.000	0.0.00
24	1.221	1.47	29.996	0.95	0.13	1.333	0.000	0.000	0.100
25	1.049	1.54	41.430	1.32	0.18	0.217	0.072	0.000	0.024
26	0.790	1.22	4.621	1.13	0.08	1.082	0.216	0.433	1.731
27	3.528	2.29	114.767	1.92	0.11	0.009	1.037	1.159	0.976
28	0.545	1.17	32.258	0.92	0.08	0.000	0.744	0.961	0.062
29	1.527	3.50	36.391	2.99	0.26	0.412	0.550	0.550	0.714

APPENDIX C: CONTINUED

			POST-ASP						
Company	Growth	Tg	MSL	*Tm*	*RSQ*	*PCT*4	*PCT*8	*PCT*12	*PCT*16
30	4.054	6.53	39.393	5.56	0.94	1.041	0.000	0.000	0.000
31	0.955	2.80	100.569	2.35	0.18	0.030	0.338	0.845	0.766
32	0.526	1.90	65.105	1.60	0.12	0.000	0.246	1.53	0.645
33	1.254	3.78	2.825	3.18	0.33	1.062	0.708	0.708	1.770
34	1.121	2.40	188.087	1.90	0.12	0.011	0.340	0.479	0.367
35	1.193	4.02	36.921	3.95	0.39	0.758	1.165	1.002	1.219
36	1.852	4.60	7.344	4.28	0.41	0.136	1.226	0.953	0.681
37	0.768	3.58	43.536	3.08	0.27	0.413	0.758	1.034	0.873
38	0.685	2.54	45.214	1.73	0.09	0.066	0.044	0.730	0.509
39	0.944	0.81	40.207	0.90	0.21	0.124	0.000	0.000	0.000
40	8.551	2.38	9.770	2.06	0.51	0.000	0.620	0.723	0.000
41	0.950	2.49	77.977	1.44	0.09	0.077	0.141	1.449	0.218
42	1.012	3.20	323.722	3.04	0.35	0.840	0.618	1.149	0.856
43	0.952	5.40	16.983	4.89	0.51	0.471	0.942	1.178	1.060
44	0.989	2.57	19.875	2.39	0.24	1.459	0.604	0.704	0.755
45	1.117	3.39	79.140	3.23	0.37	0.682	0.468	1.276	1.087
46	0.769	2.68	13.724	2.51	0.26	0.729	0.364	1.020	1.166
47	0.467	2.79	187.885	2.18	0.18	0.149	0.420	0.793	0.798
48	2.208	3.08	173.638	2.73	0.30	0.610	0.392	0.893	1.106
49	0.795	3.39	38.441	2.74	0.22	0.104	0.182	0.676	0.546
50	1.084	2.53	15.575	2.19	0.20	0.514	0.514	0.449	0.835
51	1.100	1.31	999.999	0.78	0.03	0.385	0.242	0.083	0.225
52	0.799	1.50	68.503	1.40	0.14	0.219	0.409	0.555	3.664
53	1.433	2.06	372.462	1.02	0.07	0.081	0.024	0.067	0.690
54	0.674	2.77	26.124	2.25	0.28	0.651	0.766	0.574	1.110
55	2.606	2.82	368.652	1.97	0.16	0.005	0.003	0.532	1.251
56	1.670	2.30	19.298	1.54	0.18	0.674	0.155	0.674	0.000
57	0.446	1.90	7.100	1.86	0.15	0.704	0.282	0.986	0.704
58	0.236	1.53	14.913	1.90	0.18	1.274	1.140	0.872	0.603
59	0.713	2.42	36.901	2.28	0.32	0.813	0.840	0.732	0.244
60	0.102	0.52	2.585	1.22	0.43	0.774	0.000	0.000	0.000
61	0.614	1.73	23.023	1.30	0.10	0.174	0.304	2.563	0.347
62	0.886	2.59	7.305	2.31	0.24	0.958	0.821	1.643	0.821
63	1.141	4.78	202.978	4.05	0.60	0.808	0.803	1.005	0.030
64	1.730	1.54	4.250	1.22	0.27	0.235	0.000	0.000	0.000
65	3.668	1.54	244.091	0.99	0.05	0.266	0.041	0.070	1.848
66	1.216	1.95	217.641	1.34	0.09	0.363	0.028	1.300	0.464
67	1.534	3.38	20.764	3.00	0.30	0.385	0.578	0.674	1.156
68	0.455	3.56	81.067	2.65	0.33	0.308	0.789	0.937	0.950
Average	1.387					0.576	0.593	0.812	0.779

Notes: "Growth" is the estimated growth coefficient from a single equation OLS regression of equation (7). *Tg* is the ratio of the growth coefficient to its standard error. "MSL" is calculated as the ratio of the intercept to the slope coefficient in equation (7). *RSQ* is the ratio of explained to total squares in each regression. "*PCTn*" is actual sales as a percent of the estimated MSL attained by company i at $(n-1)$ months after the date of entry. The data are wholesale shipments to drugstores during month n. In some months there were no shipments, but presumably there were sales. We did not "smooth" the data in any way; in computing the averages (*PCTn*) however we exclude from the computation any firm reporting no shipments for that month. All data and entry dates were obtained from Intercontinental Medical Statistics.

REFERENCES

Becker, Gary S., "A Theory of Competition Among Pressure Groups for Political Influence," *Quarterly Journal of Economics*, August 1983, *98*, 371–400.

Bowman, Gary Ward, "Demand and Supply of Network Television Advertising," *Bell Journal of Economics*, Spring 1976, *7*, 258–67.

Bresnahan, Timothy F., "The Demand for Advertising by Medium: Implications for the Economies of Scale in Advertising," in

P. M. Ippolito and D. T. Scheffman, eds., *Consumer Protection Economics*, Washington: Federal Trade Commission, 1986.

Brozen, Yale, "Are New FTC Policies Inhibiting Competition?," *Journal of Advertising*, March 1973, *2*, 28–31.

Clarkson, Kenneth and Muris, Timothy, *The FTC Since 1970*, New York: Cambridge University Press, 1981.

Cohen, Dorothy, "The FTC's Advertising Substantiation Program," *Journal of Marketing*, Winter 1980, *44*, 26–35.

Cox, Edward F., Fellmeth, Robert C. and Schulz, John E., *The Nader Report On the Federal Trade Commission*, New York: Richard W. Baron, 1969.

Darby, Michael R. and Karni, Edi, "Free Competition and the Optimal Amount of Fraud," *Journal of Law and Economics*, April 1973, *16*, 67–78.

Erlich, Issac and Fisher, Lawrence, "The Derived Demand for Advertising," *American Economic Review*, June 1982, *72*, 366–88.

Fama, Eugene, *Foundations of Finance*, New York: Basic, 1976.

Guerard, Collot and Niemasik, Julie, "Evolution and Evaluation of the Advertising Substantiation Program Since 1971," FTC Staff Report, 1978.

Healey, John S. and Kassarjian, Harold H., "Advertising Substantiation and Advertiser Response: A Content Analysis of Magazine Advertisements," *Journal of Marketing*, Winter 1983, *47*, 107–17.

Higgins, Richard S. and McChesney, Fred S., "Truth and Consequences: The Federal Trade Commission's Advertising Substantiation Program," *International Journal of Law and Economics*, December 1986, *6*, 151–68.

Klein, Benjamin and Leffler, Keith B., "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, *89*, 415–41.

Laband, David N., "The Durability of Informational Signals and the Content of Advertising," *Journal of Advertising*, March 1989, *18*, 13–18.

Leffler, Keith B. and Sauer, Raymond D. "The Effects of the Advertising Substantiation Program on Advertising Agencies," in P. M. Ippolito and D. T. Scheffman, eds., *Consumer Protection Economics*, Washington: Federal Trade Commission, 1986.

Peltzman, Sam, "The Effects of FTC Advertising Regulation," *Journal of Law and Economics*, December 1981, *24*, 403–48.

Posner, Richard A., "The Federal Trade Commission," *University of Chicago Law Review*, Autumn 1969, *37*, 47–89.

______, *Regulation of Advertising by the FTC*, Washington: American Enterprise Institute, 1973.

Shapiro, Carl, "Premiums for High Quality Products as Returns to Reputations," *Quarterly Journal of Economics*, November 1983, *98*, 659–79.

Stigler, George J., "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science*, Spring 1971, *2*, 3–21.

Advertising Age, various issues.

American Bar Association, *Report of the ABA Commission to Study the FTC*, Report 41, 1969.

Center for Research in Securities Prices, Daily Stock Returns Tape.

FTC Decisions, various issues.

Leading National Advertisers, *National Advertising Investments*, 1967–73.

Standard and Poor's, *Industry Studies*, 1967–73.

The Causes of Preference Reversal

By Amos Tversky, Paul Slovic, and Daniel Kahneman*

Observed preference reversal (PR) cannot be adequately explained by violations of independence, the reduction axiom, or transitivity. The primary cause of PR is the failure of procedure invariance, especially the overpricing of low-probability high-payoff bets. This result violates regret theory and generalized (nonindependent) utility models. PR and a new reversal involving time preferences are explained by scale compatibility, which implies that payoffs are weighted more heavily in pricing than in choice. (JEL 215)

Axiomatic theories of choice introduce preference as a primitive relation, which is interpreted through specific empirical procedures such as choice or pricing. Models of rational choice assume a principle of procedure invariance, which requires strategically equivalent methods of elicitation to yield the same preference order. Thus, if the decision maker prefers *A* to *B*, then the cash equivalent, or minimum selling price, of *A* should exceed that of *B*. However, there is a substantial body of evidence showing that the price ordering of risky prospects is systematically different from the choice ordering, contrary to standard theories of choice.

The effect of elicitation method on preference between gambles was first observed by Paul Slovic and Sarah Lichtenstein (1968) who found that both buying and selling prices of gambles were primarily determined by the payoffs, whereas choices between gambles (and ratings of their attractiveness) were primarily influenced by the probabilities of winning and losing. Slovic and Lichtenstein reasoned that, if the method used to elicit preferences affected the weighting of the gambles' components, it should be possible to construct pairs of gambles such that the same individual would choose one member of the pair but set a higher price for the other. Lichtenstein and Slovic (1971, 1973) demonstrated such reversals in a series of studies, one of which was conducted on the floor of the Four Queens Casino in Las Vegas with experienced gamblers playing for real money.

The preference-reversal phenomenon involves a pair of gambles of comparable expected value. One gamble (the *H* bet) offers a high probability of winning a modest sum of money; the other gamble (the *L* bet) offers a low probability of winning a relatively large amount of money. These bets were also called the *P* bet and the $ bet, respectively. For example,

H bet: 28/36 chance to win $10

L bet: 3/36 chance to win $100

When offered a choice between the two options, most subjects choose the *H* bet over the *L* bet. However, when asked to state their lowest selling price, the majority state a higher price for the *L* bet than for the *H* bet.

There have been three waves of studies of preference reversal (PR). The first included the original experiments of Paul Slovic and Sarah Lichtenstein (1968), Lichtenstein and

---

*Department of Psychology, Stanford University, Stanford, CA 94305; Decision Research, 1201 Oak Street, Eugene, OR 97401 and Department of Psychology, University of Oregon, Eugene, OR 97403; Department of Psychology, University of California, Berkeley, CA 94720. Support for this research was provided by Contract N00014-84-K-0615 from the Office of Naval Research and by grant no. 89-0064 of the Air Force Office of Scientific Research to Stanford University, by the Air Force Office of Scientific Research and by NSF grant no. SES-8712145 to Decision Research. We thank Mark Layman for his valuable assistance in the analysis of the data. The paper has benefited from discussions with Edi Karni and Eldar Shafir. We are grateful to Graham Loomes for many helpful comments.

Slovic (1971, 1973), and the study of Harold Lindman (1971). The second wave of studies, conducted during the late 1970s and early 1980s, consisted of critical replications and attempts to eliminate preference reversals by procedural variations and by increased incentives. A particularly careful replication was performed by David Grether and Charles Plott (1979) who designed a series of experiments "to discredit the psychologists' work as applied to economics" (p. 623). Grether and Plott generated a list of 13 objections and potential artifacts that would render the preference-reversal phenomenon irrelevant to economic theory. Their list included poor motivation, income effects, strategic responding, and the fact that the experimenters were psychologists (which might have led the respondents to be suspicious and behave peculiarly). Grether and Plott attempted to eliminate PR by careful experimental procedures, including a special incentive system, but the effect was hardly diminished. Further studies by Robert Hamm (1984), Werner Pommerehne, Fredrich Schneider, and Peter Zweifel (1982), and Robert Reilly (1982) led to similar conclusions. PR was also observed by John Mowen and James Gentry (1980) in both individual and group choices concerning product development, by Marc Knez and Vernon Smith (1986) who allowed their subjects to trade bets in an experimental market, and by Joyce Berg, John Dickhaut, and John O'Brien (1985) who introduced arbitrage. See also Slovic and Lichtenstein (1983).

The robustness of preference reversals led to the third wave of articles, which attempted to explain the phenomenon rather than eliminate it. Three classes of models for PR were introduced: a) nontransitive choice models were developed independently by Graham Loomes and Robert Sugden (1982, 1983), and by Peter Fishburn (1984, 1985); b) response bias models were proposed by Amos Tversky, Shmuel Sattath, and Paul Slovic (1988), and by William Goldstein and Hillel Einhorn (1987); and c) generalized utility models, which maintain transitivity but abandon the independence axiom, were applied to PR by Charles Holt (1986), and Edi Karni and Zvi Safra (1987). The three families of models respectively attribute PR to violations of transitivity, procedure invariance, or the independence axiom. Violations of these principles have been observed in other contexts, but their contributions to PR have not been assessed. The present paper provides such assessment. To accomplish this goal, we extend the traditional experimental design, introduce an ordinal payoff scheme, and develop a new diagnostic analysis to determine the causes of PR.

A. *Diagnostic Analysis.* Let C_H and C_L denote, respectively, the cash equivalent (or minimum selling price) of H and L, and $\succ$ and $\approx$ denote strict preference and indifference, respectively. (The elicitation of cash equivalents is discussed in the next section.) A standard preference reversal (PR) is said to occur when H is preferred to L, but L is priced higher than H, that is, $H \succ L$, and $C_L > C_H$. Note that $\succ$ refers to preference between options, whereas $>$ refers to the ordering of cash amounts. We assume that the two relations are consistent in the sense that $X > Y$ implies $X \succ Y$; more money is preferred to less. We now show that PR implies either intransitivity or a failure of procedure invariance, or both. Procedure invariance holds whenever the decision maker prefers a bet B to a cash amount X if and only if his or her selling price for B exceeds X; that is, $B \succ X$ iff $C_B > X$. In particular, invariance implies that $C_B = X$ iff $B \approx X$. If invariance holds, PR entails the following cycle

$$C_H \approx H \succ L \approx C_L \succ C_H,$$

where the two inequalities are implied by PR and the two equivalences follow from invariance.

Because procedure invariance is commonly taken for granted, many authors have interpreted PR as an instance of intransitivity. The finding that strategically equivalent methods of elicitation yield systematically different preferences (see, for example, Tversky et al., 1988) suggests that PR may be caused by a failure of procedure invariance, not of transitivity. Two types of discrepancy between choice and pricing could produce PR: overpricing of L and underpricing of H. *Overpricing* of L is said to occur if the decision maker prefers the price over the bet when offered a choice between them (i.e., $C_L \succ L$). *Underpricing* of H is said to occur if the decision maker prefers the bet over its price in a direct choice (i.e., $H \succ C_H$). The two other forms of a choice-pricing discrepancy—underpricing of L and overpricing of H—produce a nonstandard pattern of reversals. Note that overpricing and underpricing merely identify the sign of the discrepancy between pricing and choice; the labels do not imply that the bias resides in the pricing.

If PR is due to an intransitivity, it should be observable in binary choice data without the use of a pricing task. On the other hand, choice could be transitive if PR is due to overpricing of L or to underpricing of H. In order to find out whether PR is caused by intransitivity or by non-invariance, we extend the usual experimental design to include an option of receiving \$$X$ for sure. In this design, the decision maker provides three choices (H-L, H-X, L-X) and two assessments of price (C_H and C_L). The relevant data for diagnosing the sources of PR consists of all patterns of preferences satisfying

$$H \succ L \text{ and } C_L > X > C_H.$$

These are the patterns that exhibit PR and satisfy the condition that X lies between the two prices. Note that for any PR pattern it is possible to select a cash amount X that falls between the two prices stated by a given individual. For an experiment, however, it is more convenient to determine X in advance and to restrict the analysis to all PR patterns in which the prices stated by the subject straddle the predetermined value of X. The critical data, then, are the observed choices between X and H and between X and L. Ignoring ties, there are four possible response patterns.

(1) *Intransitivity*: $L \succ X$ and $X \succ H$, yielding $L \succ X \succ H \succ L$.

(2) *Overpricing of L (OL)*: $X \succ H$ and $X \succ L$, yielding $C_L \succ X \succ L$.

(3) *Underpricing of H (UH)*: $H \succ X$ and $L \succ X$, yielding $H \succ X \succ C_H$.

(4) *Both OL and UH*: $H \succ X$ and $X \succ L$, yielding $H \succ X > C_H$ and $C_L \succ X \succ L$.

Note that the present diagnostic procedure classifies all test patterns according to whether they violate transitivity but not invariance (pattern 1) or invariance but not transitivity (patterns 2, 3, and 4) because no test pattern can violate both. The relative frequencies of these patterns can be used to test PR models that assume invariance but not transitivity (for example, the regret theory of Graham Loomes and Robert Sugden, 1982) against models that assume transitivity but not invariance (for example, the contingent weighting model of Tversky et al., 1988). The relative incidence of violations of transitivity and invariance could be different in a more elaborate diagnostic procedure (involving more than three options) which yields patterns of preference that can violate transitivity and invariance at once. The preceding discussion is summarized as follows.

Any pattern of preferences that satisfies the test conditions ($H \succ L$ and $C_L > X > C_H$) and monetary consistency ($X > Y$ implies $X \succ Y$) obeys the following propositions:

(i) Either transitivity or invariance ($B \succ X$ iff $C_B > X$) is violated.

(ii) Intransitivity holds iff $L \succ X \succ H$, yielding pattern (1).

(iii) Overpricing of L occurs iff $X \succ L$, yielding patterns (2) or (4).

(iv) Underpricing of H occurs iff $H \succ X$, yielding patterns (3) or (4).

This analysis provides a simple method for the diagnosis of all preference patterns satisfying the test condition according to the implied cause of PR. The applications of this method to experimental data are described in Studies 1 and 2.

B. *Elicitation Procedures.* The preceding analysis provides an effective method for diagnosing PR, assuming preferences and selling prices are observable. Such data are routinely elicited in experiments and surveys, and are normally taken at face value—provided the results are reproducible and the respondents have no reason to conceal or misrepresent their preferences. Nevertheless, students of choice favor elicitation procedures that are incentive compatible. Gordon Becker, Morris DeGroot, and Jacob

Marschak (1964) devised such a scheme for the elicitation of cash equivalents, which has often been used in PR studies. In this procedure, called the BDM scheme, the subject states a selling price for a gamble. An offer is then generated by some random process. The subject receives the offer if it exceeds the stated selling price, and plays the gamble if the stated price exceeds the offer.

Recently, Holt (1986), and Karni and Safra (1987) pointed out that the justification of the BDM scheme and of similar procedures assumes the independence axiom of expected utility theory. If this axiom fails, the stated price of a bet is no longer equal to its cash equivalent, hence the presence of PR in some experiments can be attributed to violations of independence. In particular, Holt argued that the usual payoff schemes in which the subject plays only one bet, selected at random, assume independence among the subject's responses. This argument, however, invokes the independence axiom in a sequential setup where no systematic violations of this axiom have been observed. Moreover, Tversky and Daniel Kahneman (1986) showed that the common violations of independence and substitution generally disappear when the choice is presented in a sequential form. This finding undermines the empirical basis of Holt's argument. Karni and Safra analyzed the actual lottery induced by the BDM scheme and showed that a generalized utility model with a nonlinear (probability) weighting function can explain the observed discrepancy between choice and stated prices in this scheme. According to this interpretation, PR data do not reflect inconsistent preferences because the choices and the pricing do not involve the same bets. Consequently, PR can be explained by a model that satisfies transitivity and invariance, but not independence. In a closely related development, Uzi Segal (1988) suggested that PR may be produced by a failure of the reduction axiom, which reduces a two-stage lottery to the equivalent one-stage lottery. The preceding accounts, however, cannot explain the basic finding that the same pattern of PR obtained under the BDM scheme (for example, Grether and Plott, 1979) is equally prevalent in experiments that do not use any incentive compatible elicitation scheme (for example, Lichtenstein and Slovic, 1971) and therefore do not invoke the independence or the reduction axioms. If PR is caused by a failure of these axioms, the frequency of reversals should increase when the BDM scheme is used, contrary to fact.

As noted by Karni and Safra (1987), there may be no incentive compatible scheme for the elicitation of selling prices that does not rely on the independence axiom. To demonstrate preference reversals, however, it is not necessary to elicit the actual selling prices; it is sufficient to establish their *order*. The ordering of selling prices can be obtained under weaker conditions, as demonstrated by the following incentive compatible procedure, which we call the ordinal payoff scheme. The subject is first presented with each bet separately and asked to state its lowest selling price. The subject is then presented with pairs of bets and asked to select the preferred prospect in each pair. Subjects know that one of these pairs will be selected at random at the end of the session, and that a random device will determine whether they play the bet they preferred in the choice task or the bet they priced higher. Because the prices are merely used to order the bets within each pair, choice and pricing are strategically equivalent. A reversal then amounts to an inconsistency in which the subject expresses in the choice task a preference for playing the H bet, and in the pricing task a preference for playing the L bet. To rationalize this pattern within a generalized (nonindependent) utility model, it is necessary to assume that subjects prefer an even chance to play *either* H or L over the option of playing H, *and* over the option of playing L. Such a mixed strategy, however, could explain random reversals, but not systematic ones. The presence of systematic reversals in the ordinal payoff scheme, therefore, cannot be explained by violations of independence or of the reduction axiom.

I. Study 1 (Monetary Bets)

The participants in the studies reported in this article (unless specified otherwise) were

TABLE 1—THE MONETARY BETS USED IN STUDY 1.[a]

Set	Triple	1	2	3	4	5	6
	H	(0.97,4)	(0.81,2)	(0.94,3)	(0.89,4)	(0.94,2.5)	(0.92,2)
I	L	(0.31,16)	(0.19,9)	(0.50,6.5)	(0.11,40)	(0.39,8.5)	(0.50,5)
	X	(3.85)	(1.50)	(2.75)	(3.25)	(2.40)	(1.85)
	H	(0.97,100)	(0.81,50)	(0.94,75)	(0.89,100)	(0.94,65)	(0.92,50)
II	L	(0.31,400)	(0.19,225)	(0.50,160)	(0.11,1000)	(0.39,210)	(0.50,125)
	X	(95)	(40)	(70)	(80)	(60)	(45)
	H	(0.78,10)	(0.69,7)	(0.86,3)	(0.94,4)	(0.92,12)	(0.89,11)
III	L	(0.08,100)	(0.17,40)	(0.19,13)	(0.03,150)	(0.06,175)	(0.08,135)
	X	(8)	(5.50)	(2.75)	(3.75)	(10)	(9)

[a] The pair (P, X) denotes a bet that offers a probability P to win $\$X$, and nothing otherwise, and (X) denotes $\$X$ for sure. The probabilities were expressed as multiples of 1/36.

recruited by ads in the University of Oregon newspaper. The number of men and women was roughly the same and their median age was 22 years. They were run in a class setting and were paid for participation. One hundred and ninety-eight individuals (called the main group) participated in the main part of Study 1.

We constructed 18 triples that consisted of a high probability bet (H), a low probability bet (L), and a cash amount (X), with roughly the same expected value. The 18 triples were divided into 3 sets of 6 triples each, displayed in Table 1. Set I included relatively small bets like those used in previous studies, without negative outcomes. Set II was obtained by multiplying all payoffs in Set I by a factor of 25. Set III was a mixture of small and large bets including a few long shots. The options were presented in written form. All probabilities were expressed as multiples of 1/36 (for example, 11/36 to win \$16.00), and interpreted with reference to a roulette wheel with 36 numbered sectors. Within each set, the respondents first assessed the cash equivalent for each of the 12 bets and then chose between the options of each triple: H vs. L, H vs. X, L vs. X. The order of the options within each set was randomized. The order of the sets was I, II, III for half of the participants, and III, II, I for the others. As part of the general instructions, the participants were asked to imagine that one pair of bets would be selected at random and they would have a chance to play either the bet they had chosen or the bet for which they had stated a higher cash equivalent.

To investigate the effect of monetary incentives, we ran an additional group of subjects whose payoffs were contingent on their responses. These subjects ($N = 179$) evaluated only Set I. They received the same instructions as the main group, but were told that 15 percent of the participants, selected at random, would actually play one bet determined as follows. A pair of bets would be selected at random and the chosen subjects would play the bet that they had favored either in the choice or in the pricing task. Because each subject plays at most one bet, this procedure defines a compound lottery in which the probability of winning a prize is considerably smaller than in the original bet. Exactly the same compound lottery, however, is induced by the choice and by the pricing. Assuming invariance, therefore, there is no reason for reversals in the ordinal payoff scheme.

A. *Results*. Table 2 presents, for each (H, L) pair, the percentage of subjects who chose H over L ($H \succ L$), the percentage of subjects who priced H above $L (C_H > C_L)$, and the percentage of subjects who exhibited PR($H \succ L$ and $C_L > C_H$). Only the subjects ($N = 179$) who had no missing data were included in this analysis. Because no indif-

TABLE 2—PERCENTAGE OF PREFERENCES FOR ALL SUBJECTS IN STUDY 1

Set	Triple	1	2	3	4	5	6	Mean
	$H \succ L$	83	68	71	71	73	62	71
I	$C_H > C_L$	26	22	30	33	17	14	24
	PR	59	53	45	41	59	49	51
	$H \succ L$	91	86	77	84	82	70	82
II	$C_H > C_L$	54	48	46	47	48	32	46
	PR	38	44	36	41	40	44	41
	$H \succ L$	81	68	74	74	81	79	76
III	$C_H > C_L$	38	21	39	38	58	46	40
	PR	48	48	42	39	29	36	40
	$H \succ L$	67	75	62	67	62	56	65
I_s	$C_H > C_L$	44	18	22	44	12	9	25
	PR	32	60	51	36	50	49	46

ference was allowed in choice, we excluded tied prices from the analysis. Their inclusion, of course, would have inflated the PR rate. The $\mathrm{I}_\$$ data refer to the subjects who evaluated Set I with monetary incentives.

The PR phenomenon is clearly evident in Table 2. Overall, H is chosen over L in 74 percent of the cases, but C_H exceeds C_L in only 34 percent of the cases. Nearly half of the response patterns are inconsistent and the standard reversals (45 percent) are overwhelmingly more frequent than the nonstandard ones (4 percent). Table 2 also shows that there are no major differences among three sets of gambles that span a wide range of payoffs. Furthermore, the use of monetary incentives in the $\mathrm{I}_\$$ condition had no systematic effect on the prevalence of PR, in agreement with previous studies (see, for example, Slovic and Lichtenstein, 1983). The overall percentage of PR with real payoffs is 46 percent (Set $\mathrm{I}_\$$) compared to an overall rate of 44 percent without payoffs (Sets I, II, III).

The hypothesis that subjects prefer mixtures was tested by presenting 42 Stanford students with the six pairs of H and L bets from Set I. They were told that one-third of the participants, selected at random, would play one of their chosen bets. For each pair, they were given three options: (i) select the H bet, (ii) select the L bet, and (iii) let the experimenter select between H and L at random. None of the responses favored the mixed strategy.

The results of Study 1 cannot be explained as a failure of the independence axiom, or of the reduction axiom, for two reasons. First, this account assumes that the subjects favor a mixed strategy, contrary to the responses of the Stanford students. Second, even if—contrary to these data—we assume that subjects adopt a mixed strategy, the standard and the nonstandard reversals should have been equally frequent. We conclude that PR is not an elicitation artifact caused by nonindependence; hence it must violate either transitivity or procedure invariance.

Having established the validity of PR, we turn now to the diagnostic analysis described earlier. Unlike the preceding discussion that was based only on the price ordering, the following analysis assumes that the stated prices C_H and C_L are the actual cash equivalents of H and L. We shall address this assumption in the next section. The diagnostic analysis is based on the data of all subjects from the main group who completed all tasks. About half of the PR patterns in Sets I, II, and III met the test condition $C_L > X > C_H$, yielding 620 individual patterns of preference that satisfied

$$H \succ L \text{ and } C_L > X > C_H.$$

TABLE 3—DISTRIBUTION OF RESPONSE PATTERNS IN STUDY 1

Pattern	N	Percent	Diagnosis
(1) $L \succ X$, $X \succ H$	62	10.0	Intransitivity
(2) $X \succ L$, $X \succ H$	406	65.5	Overpricing of L (OL)
(3) $L \succ X$, $H \succ X$	38	6.1	Underpricing of H (UH)
(4) $H \succ X$, $X \succ L$	114	18.4	Both OL and UH
Total	620	100	

These data were classified into four types according to the observed relation of X to L and to H, and the distribution of the four types is presented in Table 3. The data show that only 10 percent of the preference patterns are intransitive and 90 percent exhibit non-invariance. Among the non-invariant patterns, the overpricing of L is much more common than the underpricing of H. Further evidence for this conclusion comes from a study by Raphael Bostic, Richard Herrnstein, and Duncan Luce (1990) who used a series of successive choices to determine, for each subject, a cash amount X that is indifferent to a given bet. They found that this cash amount was roughly the same as the stated price for H bets and substantially below it for L bets. That is, $X = C_H$ and $X < C_L$, as entailed by the overpricing of L.

B. *Discussion*. The high incidence of preference reversal (45 percent) obtained using the ordinal payoff scheme indicates that PR cannot be explained by a generalized utility model, as proposed by Holt (1986), by Karni and Safra (1987), and by Segal (1988). Recall that these authors provided no direct evidence for this proposal; they only showed that PR could be consistent with transitivity and invariance if independence (or the reduction axiom) is discarded. The data refute this interpretation. The present analysis demonstrates that abandoning independence or the reduction axiom is neither necessary nor sufficient to account for PR. The results summarized in Table 3 show that the PR observed in Study 1 are produced primarily by the failure of procedure invariance. The finding that 90 percent of test patterns violated this condition rules out all models of PR that assume procedure invariance, including regret theory (Loomes and Sugden, 1982, 1983). Because this theory attributes all PR to intransitivity, it entails that 100 percent of the test patterns in Table 3 should be intransitive. The fact that only 10 percent of the test patterns conformed to this prediction indicates that regret theory does not provide an adequate explanation of PR.

This conclusion does not imply that intransitivity plays no role in PR. Loomes, Chris Starmer, and Sugden (1989) have demonstrated that the cycle $H \succ L \succ X \succ H$, implied by regret theory, is significantly more frequent than the opposite cycle $L \succ H \succ X \succ L$. The present study confirmed this result; the observed frequencies of the two cycles were 132 and 44, respectively (see Appendix 1). It is noteworthy that the incidence of intransitivity reported by these investigators (15–20 percent) is higher than that observed in the present study, but the rates of PR are roughly the same. Although the difference in the incidence of intransitivity does not affect our main conclusion regarding the prevalence of non-invariance, it raises the question as to whether the present methodology underestimates the contribution of intransitivity of PR. Because the ordinal payoff scheme requires, in effect, the ranking of all bets, it might induce intransitive subjects to behave transitively. But if our procedure reduces the number of cycles, it should also reduce the incidence of PR. In fact, the PR rate in the present study is slightly higher than in other studies. Moreover, the finding that the rate of PR is generally insensitive to the payoff scheme makes this possibility even less likely. Finally, Appendix 1 shows that the percentage of cycles that satisfy the test condition (10 percent) is higher than the percentage of cycles among PR patterns that do not satisfy the test condition (5.2 percent). The available evidence, therefore, does not support the hypothesis that the present diagnostic procedure underestimates intransitivity, although this possibility cannot be ruled out.

It could be argued that the subjects in Study 1 did not state their actual selling prices but rather some monotone transformation of these prices because only their order—not the actual values—is relevant to the present payoff scheme. The subjects, however, had no strategic or any other reason to transform their prices. Moreover, the present procedure discourages such transformation because the subject has to evaluate each bet separately, without knowing how it will be paired. Stating the cash equivalent of each bet, as requested, is obviously the simplest way to achieve a ranking without having to remember the responses to previous bets. Indeed, the standard deviations (across subjects) of observed prices in Study 1 did not exceed those obtained from comparable bets in studies that used the BDM scheme (for example, Bostic, Herrnstein, and Luce, 1990), contrary to what might be expected if our subjects reported monotone transformations of their selling prices. As in other studies, the standard deviations were considerably smaller for *H* bets than for *L* bets.

In contrast to the present design where each bet was priced separately and the subject did not know which bets would be paired, James Cox and Seth Epstein (1989) employed a different procedure in which the subject priced each pair of *H* and *L* bets concurrently, and then played the higher-priced bet. This procedure allows the subject to generate a price ordering by comparing the two bets without performing the more difficult task of assessing their actual selling prices. Indeed, Cox and Epstein acknowledged (on the basis of the subjects' responses) that their procedure did not elicit proper cash equivalents. Because the bets were framed differently in choice and in pricing (by translating all outcomes), many response patterns were inconsistent (32 percent), but the reversals were random rather than systematic.

C. *Scale Compatibility*. Why do people overprice the low-probability high-payoff bets? Following Tversky et al. (1988), we interpret this finding as an instance of a more general principle of compatibility: the weight of any aspect (for example, probability, payoff) of an object of evaluation is enhanced by compatibility with the response (for example, choice, pricing). The effect of scale compatibility is illustrated in a study by Slovic, Dale Griffin, and Tversky (1990). Participants predicted the 1987 market value of twelve companies (taken from *Business Week* top 100) on the basis of their 1986 market value (in billions of dollars), and their rank (among the top 100) with respect to 1987 profits. One-half of the subjects predicted 1987 market value in billions of dollars, whereas the other half predicted the company's rank with respect to its 1987 market value. As implied by compatibility, each predictor was given more weight when the criterion was expressed on the same scale (for example, money, rank). As a consequence, the relative weight of the 1986 market value was twice as high for those who predicted in dollars than for those who predicted the corresponding rank. This effect produced many reversals of judgment in which one company was ranked above another but the order of their predicted value was reversed.

Because the selling price of a bet is expressed in dollars, compatibility entails that the payoffs, which are expressed in the same units, will be weighted more heavily in pricing than in choice. (Since the payoffs are much larger in the *L* bets than in the *H* bets, the major consequence of a compatibility bias is overpricing of *L* bets.) To accommodate violations of procedure invariance, Tversky et al. (1988) developed a family of models in which the weighting of attributes is contingent on the method of elicitation. The simplest model of this type, called the contingent weighting model, assumes that the bet $B = (P, X)$ is chosen over $B' = (P', X')$ iff

$$\log P + \alpha \log X > \log P' + \alpha \log X'.$$

and that B is priced above B' iff

$$\log P + \beta \log X > \log P' + \beta \log X'.$$

This loglinear form represents the ordering of bets according to both choice and pricing by a multiplicative probability-value model with a power function for gains. In this representation, procedure invariance entails $\beta = \alpha$, whereas compatibility implies $\beta > \alpha$. Slovic et al. (1990) applied this model to the choice and the price ordering from Study 1 and assessed α and β separately for each subject. In accord with compatibility, β exceeded α for 87 percent of the subjects ($N = 179$) indicating greater risk aversion in choice than in pricing.

The role of compatibility in PR was investigated in two additional studies reported by Slovic et al. (1990). In the first study, subjects were presented with six pairs of H and L bets. Three pairs involved monetary payoffs as in Study 1, and three pairs involved nonmonetary outcomes, such as a one-week pass for all movie theaters in town, or a dinner for two at a good restaurant. If preference reversals are due primarily to the compatibility of prices and payoffs, which are both expressed in dollars, their incidence should be substantially reduced by the use of nonmonetary outcomes. This prediction was confirmed: the percentage of choice of H over L was roughly the same in the monetary (63 percent) and the nonmonetary bets (66 percent), but the percentage of cases in which $C_L > C_H$ was substantially greater in the former (54 percent) than in the latter (33 percent), and the overall incidence of PR decreased significantly from 41 percent to 24 percent.

A second study by Slovic et al. (1990) investigated reversals in a matching task. This experiment employed the 12 pairs of H and L bets from Sets I and II, described in Table 1. In each pair, one value—either a probability or a payoff—was missing, and the subjects were asked to set the missing value so that they would be indifferent between the two bets. Consider, for example, the bets $H = (0.92, \$50)$ and $L = (0.50, \$125)$ from Set II, Table 1. If we replace the 0.5 probability in L by a question mark, the subject is asked in effect "what chance to win $125 is equally attractive as a 0.92 chance to win $50?," The value set by the subject implies a preference between the original bets. If the value exceeds 0.50, we infer that the subject prefers H to L, and if the value is less than 0.50, we reach the opposite conclusion. Using all four components as missing values, the authors obtained the preferences inferred from matching either the probability or the payoff of each bet. Because compatibility applies to matching, not only to pricing, the attribute on which that match is made is expected to be weighted more heavily than the other attribute. As a consequence, the inferred percentage of preference for H over L should be higher for probability matches than for payoff matches. Indeed, the overall percentage of preference for H over L derived from probability matches (73 percent) was significantly higher than that derived from payoff matches (49 percent).

The matching data suggest an additional factor, besides scale compatibility, that might contribute to PR. In an extensive comparison of choice and matching, Tversky et al. (1988) have shown that the more prominent attribute of options looms larger in choice than in matching. Thus, the choice order is more lexicographic than that induced by matching. If probability is perceived by subjects as more important than payoffs, as suggested by the finding that the rating of bets is dominated by probability (Goldstein and Einhorn, 1987; Slovic and Lichtenstein, 1968), then the prominence effect may contribute to PR over and above the effect of scale compatibility. This could help explain the robustness of PR, as well as the similarity between the choice data and the preferences derived from probability matches. Note that in pricing and payoff matches, both compatibility and prominence enhance the PR effect, whereas in probability matches, they operate in opposite directions (see, Slovic et al., 1990).

II. Study 2 (Time Preferences)

According to compatibility, reversals of preference are not restricted to risky prospects; they should also occur in other situations in which decision makers both choose and price options with a monetary component. In the next study, we test this

TABLE 4—THE OPTIONS USED IN STUDY 2 AND THE RESPECTIVE PERCENTAGE OF PREFERENCES. THE PAIR (X, T) DENOTES THE OPTION OF RECEIVING $\$X$, T YEARS FROM NOW, AND (X) DENOTES AN IMMEDIATE CASH PAYMENT

Triple	1	2	3	4	Mean
S	(1600,1.5)	(1600,1.5)	(2500,5)	(1525,0.5)	
L	(2500,5)	(3550,10)	(3550,10)	(1900,2.5)	
X	(1250)	(1250)	(1250)	(1350)	
$S \succ L$	57	72	83	83	74
$C_S > C_L$	12	19	29	40	25
PR	49	56	57	46	52

prediction in the context of time preferences. Consider a delayed payment of the form (X, T) that offers a payment of $\$X$, T years from now. Following the design of Study 1, we constructed four triples of options that consisted of a long-term prospect L (for example, \$2500, 5 years from now), a short-term prospect S (for example, \$1600, 1.5 years from now), and an immediate payment X (for example, \$1250 now).

One hundred and sixty-nine subjects participated in the study. One-half of the subjects first chose between the three pairs of options (S vs. L, S vs. X, L vs. X) in each triple. The order of the pairs was randomized across triples. Following the choice, these subjects priced each of the S and L options by stating "the smallest immediate cash payment for which they would be willing to exchange the delayed payment." The other subjects performed the choice and pricing tasks in the opposite order. There were no systematic differences between the groups, so their data were combined.

The upper part of Table 4 presents the four triples of options employed in this study. The lower part of the table includes, for each triples, the percentage of subjects who chose S over L ($S \succ L$), the percentage of subjects who priced S above L ($C_S > C_L$), and the percentage of PR patterns ($S \succ L$ and $C_L > C_S$). Table 4 reveals a massive amount of PR. Overall, the short-term option (S) was chosen over the long-term option (L) 74 percent of the time, but was priced higher only 25 percent of the time, yielding more than 50 percent PR patterns.

Note that in the pricing task each option is evaluated singly whereas choice involves a comparison between options. The observed reversals, therefore, are consistent with an alternative hypothesis that payoffs are weighted more heavily in a singular than in a comparative evaluation. To test this hypothesis, we replicated the study in a new group of 184 subjects, with one change. Instead of pricing the options, the subjects were asked to rate the attractiveness of each option on a scale from 0 (not at all attractive) to 20 (extremely attractive). If PR is controlled, in part at least, by the nature of the task (singular vs. comparative), rating and pricing should yield similar results. On the other hand, if PR is produced primarily by the compatibility between payoffs and prices, rating and choice should yield similar results. The data reveal no discrepancy between choice and rating. Overall, S was chosen over L 75 percent of the time (as in the original study) and the rating of S exceeded the rating of L in 76 percent of the cases, in accord with procedure invariance. Only 11 percent of the patterns exhibited PR.

We also applied the diagnostic analysis to the results of Study 2. Recall that the relevant test data consisted of all patterns of the form

$$S \succ L \text{ and } C_L > X > C_S.$$

These are the patterns that exhibit preference reversal, and the immediate payment of $\$X$ falls between the two cash equivalents. These data were classified into four types according to the observed relation of X to L and to S. The distribution of the four types is presented in Table 5 along with the appropriate diagnoses. The table shows that the

TABLE 5—DISTRIBUTION OF RESPONSE PATTERNS IN STUDY 2.

Pattern	N	Percent	Diagnosis
(1) $L \succ X$, $X \succ S$	16	15.4	Intransitivity
(2) $X \succ L$, $X \succ S$	57	54.8	Overpricing of L (OL)
(3) $L \succ X$, $S \succ X$	12	11.5	Underpricing of S (US)
(4) $S \succ X$, $X \succ L$	19	18.3	Both OL and US
Total	104	100	

major determinant of PR in delayed payments is the overpricing of the long-term prospect (L), and that the underpricing of the short-term prospect (S) and the intransitivity of $\succ$ play a relatively minor role (see also Appendix 2). Evidently, payoffs are weighted more heavily in pricing than in choice.

III. General Discussion

The preference-reversal phenomenon has been demonstrated in numerous studies but its causes have not been established. The present article attempts to determine these causes by comparing several accounts that attribute PR to violations of transitivity, procedure invariance, independence, or the reduction axiom. To assess the relative contributions of these effects we extend the standard experimental design, employ an ordinal payoff scheme, and introduce a new diagnostic analysis. The results appear quite clear.

First, PR cannot be attributed to violations of the independence or the reduction axioms and therefore cannot be rationalized by a generalized (nonindependent) utility model, as proposed by Holt (1986), Karni and Safra (1987), and Segal (1988). Note that the basic PR pattern does not involve these axioms, which are only needed to justify payoff procedures such as the BDM scheme. If PR were caused by nonindependence, or by a failure of the reduction axiom, it should be observed in the BDM scheme but not in the present ordinal scheme or in experiments where the payoffs are not contingent on the subject's responses. The finding that PR is no more frequent in the BDM scheme, which assumes independence, than in elicitation procedures that do not assume this axiom, rules out nonindependence as a viable explanation of PR. Evidently, the effect of response mode (for example, choice, pricing) cannot be explained by the payoff scheme.

If PR does not represent a violation of independence or of the reduction axiom, it must be due either to intransitivity or to a failure of procedure invariance. The diagnostic procedure indicates that only 10 percent of PR patterns are intransitive while the remaining 90 percent represent non-invariance. Because Fishburn's (1984, 1985) SSB model and Loomes and Sugden's (1982, 1983) regret theory assume procedure invariance and attribute all PR to intransitivity, they do not provide an adequate account of preference reversals. It is also noteworthy that regret theory is consistent with the predominant direction of the (relatively small percentage of) intransitive triples observed in Study 1, but is not consistent with the pronounced intransitivity of choices between gambles demonstrated by Tversky (1969).

The overpricing of the low-probability high-payoffs bets emerges from the present analysis as the major cause of PR. The results of Raphael Bostic, Richard Herrnstein, and Duncan Luce (1990) provide further support for this conclusion. We have interpreted the overpricing of long shots as an effect of scale compatibility: because the prices and the payoffs are expressed in the same units, payoffs are weighted more heavily in pricing than in choice. Compatibility also implies new preference reversals in other domains, such as time preferences. Study 2 shows that when faced with a choice between delayed payments, people frequently select the short-term prospect but assign a higher cash equivalent to the long-term prospect. Other consequences of compatibility have been confirmed in other studies showing (i) that the prevalence of PR is substantially reduced by using nonmonetary payoffs, and (ii) that the matched attribute is weighted more heavily than the other attribute in a matching task. Further evidence on the role of compatibility in PR, based on attention and anchoring data, is presented by David Schkade and Eric Johnson (1989). We conclude that scale compatibility explains, at least in part, the primary cause of PR, namely the overpricing of *L* bets. It should be noted, however, that compatibility does not explain several secondary effects commonly observed in PR studies, including the occurrence of systematic intransitivities and the occasional underpricing of *H* bets.

Experimental studies of choice challenge the classical analysis of decision making. Violations of the independence axiom of expected utility theory have attracted much attention. In addition, a growing body of empirical evidence questions the assumption of invariance, which is essential to the theory of rational choice. In particular, alternative framings of the same options (for example, in terms of gains vs. losses, or in terms of survival vs. mortality) produce inconsistent preferences (Tversky and Kahneman, 1986), and alternative elicitation procedures (for example, choice, pricing) give rise to reversal of preferences (Tversky, et al. 1988). These results confirm the conclusion of Grether and Plott:

> Taken at face value the data are simply inconsistent with preference theory and have broad implications about research priorities within economics. The inconsistency is deeper than the mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind the simplest of human choices and that the uniformities in human choice behavior which lie behind market behavior may result from principles which are of a completely different sort from those generally accepted.
>
> [1979, p. 623]

Indeed, the failures of description invariance (framing effects) and procedure invariance (elicitation effects) pose a greater problem for rational choice models than the failure of specific axioms, such as independence or transitivity, and they demand descriptive models of much greater complexity. Violations of description invariance require an explicit treatment of the framing process, which precedes the evaluation of prospects (Kahneman and Tversky, 1979). Violations of procedure invariance require context-dependent models (for example, Tversky et al., 1988) in which the weighting of attributes is contingent on the method of elicitation. These developments highlight the discrepancy between the normative and the descriptive approaches to decision making, which many choice theorists (see Mark Machina, 1987) have tried to reconcile. Because invariance—unlike independence or even transitivity—is normatively unassailable and descriptively incorrect, it does not seem possible to construct a theory of choice that is both normatively acceptable and descriptively adequate.

APPENDIX 1–DISTRIBUTION OF RESPONSE PATTERNS IN STUDY 1

	1	2	3	4	5	6	
	$C_L > X > C_H$	$C_H > X > C_L$	$C_L > C_H > X$	$C_H > C_L > X$	$X > C_L > C_H$	$X > C_H > C_L$	Total
1. $X \succ H$, $H \succ L$, $X \succ L$	406	142	154	2	203	427	1334
2. $H \succ X$, $L \succ H$, $L \succ X$	81	8	61	0	13	25	188
3. $X \succ H$, $H \succ L$, $L \succ X$	62	8	15	0	25	22	132
4. $H \succ X$, $L \succ H$, $X \succ L$	20	4	7	0	2	11	44
5. $X \succ H$, $L \succ H$, $X \succ L$	81	10	17	1	32	26	167
6. $H \succ X$, $H \succ L$, $L \succ X$	38	11	33	0	13	11	106
7. $X \succ H$, $L \succ H$, $L \succ X$	158	15	54	1	52	19	299
8. $H \succ X$, $H \succ L$, $X \succ L$	114	114	89	2	22	128	469
Total	960	312	430	6	362	669	2739

APPENDIX 2–DISTRIBUTION OF RESPONSE PATTERNS IN STUDY 2

	1	2	3	4	5	6	
	$C_L > X > C_S$	$C_S > X > C_L$	$C_L > C_S > X$	$C_S > C_L > X$	$X > C_L > C_S$	$X > C_S > C_L$	Total
1. $X \succ S, S \succ L, X \succ L$	57	14	63	5	32	65	236
2. $S \succ X, L \succ S, L \succ X$	19	0	59	1	4	1	84
3. $X \succ S, S \succ L, L \succ X$	16	1	10	1	4	1	33
4. $S \succ X, L \succ S, X \succ L$	1	2	7	0	4	4	18
5. $X \succ S, L \succ S, X \succ L$	14	3	9	0	2	3	31
6. $S \succ X, S \succ L, L \succ X$	12	2	24	3	5	1	47
7. $X \succ S, L \succ S, L \succ X$	9	1	12	1	2	2	27
8. $S \succ X, S \succ L, X \succ L$	19	11	42	8	12	18	110
Total	147	34	226	19	65	95	586

REFERENCES

Becker, Gordon M., DeGroot, Morris H. and Marschak, Jacob, "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, July 1964, *9*, 226–32.

Berg, Joyce E., Dickhaut, John W. and O'Brien, John R., "Preference Reversal and Arbitrage," in V. L. Smith, ed., *Research in Experimental Economics*, Vol. 3, Greenwich: JAI Press, 1985.

Bostic, Raphael, Herrnstein, Richard J. and Luce, R. Duncan, "The Effect on the Preference-Reversal Phenomenon of Using Choice Indifferences," *Journal of Economic Behavior and Organization*, 1990, in press.

Cox, James C. and Epstein, Seth, "Preference Reversals Without the Independence Axiom," *American Economic Review*, June 1989, *79*, 408–26.

Fishburn, Peter C., "SSB Utility Theory: An Economic Perspective," *Mathematical Social Sciences*, August 1984, *8*, 63–94.

______, "Nontransitive Preference Theory and the Preference Reversal Phenomenon," *Rivista Internazionale di Scienze Economiche e Commerciali*, January 1985, *32*, 39–50.

Goldstein, William M. and Einhorn, Hillel J., "Expression Theory and the Preference Reversal Phenomena," *Psychological Review*, April 1987, *94*, 236–54.

Grether, David M. and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *69*, 623–38.

Hamm, Robert M., "The Conditions of Occurrence of the Preference Reversal Phenomenon," Technical Report, Center for Research on Judgment and Policy, University of Colorado, 1984.

Holt, Charles A., "Preference Reversals and the Independence Axiom," *American Economic Review*, June 1986, *76*, 508–15.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

Karni, Edi and Safra, Zvi, "'Preference Reversal' and the Observability of Preferences by Experimental Methods," *Econometrica*, May 1987, *55*, 675–85.

Knez, Marc and Smith, Vernon L., "Hypothetical Valuations and Preference Reversals in the Context of Asset Trading," unpublished manuscript, University of Arizona, 1986.

Lichtenstein, Sarah and Slovic, Paul, "Reversals of Preference between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, January 1971, *89*, 46–55.

______, "Response-Induced Reversals of Preference in Gambling: An Extended Replication in Las Vegas," *Journal of Experimental Psychology*, November 1973, *101*, 16–20.

Lindman, Harold R., "Inconsistent Preferences Among Gambles," *Journal of Experimental Psychology*, May 1971, *89*, 390–97.

Loomes, Graham and Sugden, Robert, "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *Economic Journal*, December 1982, *92*, 805–24.

______ **and** ______, "A Rationale for Preference Reversal," *American Economic Review*, June 1983, *73*, 428–32.

______, **Starmer, Chris and Sugden, Robert,** "Preference Reversal: Information-Processing Effect or Rational Nontransitive Choice," *Economic Journal* (Conference Supplement), 1989, *99*, 140–51.

Machina, Mark J., "Choice Under Uncertainty: Problems Solved and Unsolved," *Economic Perspectives*, Summer 1987, *1*, 121–54.

Mowen, John C. and Gentry, James W., "Investigation of the Preference-Reversal Phenomenon in a New Product Introduction Task," *Journal of Applied Psychology*, December 1980, *65*, 715–22.

Pommerehne, Werner W., Schneider, Fredrich and Zweifel, Peter, "Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination," *American Economic Review*, June 1982, *72*, 569–74.

Reilly, Robert J., "Preference Reversal: Further Evidence and Some Suggested Modifications in Experimental Design," *American Economic Review*, June 1982, *72*, 576–84.

Schkade, David A. and Johnson, Eric J., "Cognitive Processes in Preference Reversals," *Organizational Behavior and Human Performance*, June, 1989, *44*, 203–31.

Segal, Uzi, "Does the Preference Reversal Phenomenon Necessarily Contradict the Independence Axiom?," *American Economic Review*, March 1988, *78*, 233–36.

Slovic, Paul, Griffin, Dale and Tversky, Amos, "Compatibility Effects in Judgment and Choice," in R. M . Hogarth, ed., *Insights in Decision Making: Theory and Applications*, Chicago: University of Chicago Press, 1990.

______, **and Lichtenstein, Sarah,** "The Relative Importance of Probabilities and Payoffs in Risk-Taking," *Journal of Experimental Psychology Monograph Supplement*, November 1968, Part 2, *78*, 1–18.

______, "Preference Reversals: A Broader Perspective," *American Economic Review*, September 1983, *73*, 596–605.

Tversky, Amos, "Intransitivity of Preferences," *Psychological Review*, January 1969, *76*, 31–48.

______, **and Kahneman, Daniel,** "Rational Choice and the Framing of Decisions," *Journal of Business*, October 1986, *59*, 251–78.

______, **Sattath, Shmuel and Slovic, Paul,** "Contingent Weighting in Judgment and Choice," *Psychological Review*, July 1988, *95*, 371–84.

Selection Criteria in Coordination Games: Some Experimental Results

By RUSSELL W. COOPER, DOUGLAS V. DEJONG, ROBERT FORSYTHE, AND THOMAS W. ROSS*

A weakness of the Nash equilibrium concept for noncooperative games is that it may not generate a unique outcome. In this case, it must be augmented by a hypothesis refining the beliefs of players about the strategies selected by their opponents. As beliefs are inherently unobservable, one means of evaluating these hypotheses is through inferences based upon the observed play of participants in experimental games.

We study a class of symmetric, simultaneous move, complete information games called *coordination games*. This term refers to games which exhibit multiple Nash equilibria which are Pareto-rankable.[1] That is, all players are better off in one equilibrium relative to another yet may be unable to explicitly coordinate their strategies to achieve the preferred outcome. When this occurs, a coordination failure arises.

These games characterize strategic interactions in a large number of settings. Schelling (1960, 1978) provides an overview of social environments in which coordination games naturally arise. Coordination games are at the heart of a number of recent models in industrial organization and macroeconomics. These include models of network externalities (Michael Katz and Carl Shapiro, 1985), product warranties under bilateral moral hazard (Cooper-Ross, 1985), team production (John Bryant, 1983); and macroeconomic models of imperfect competition (Walter Heller, 1986 and Nobuhiro Kiyotaki, 1988) and search (Peter Diamond, 1982).[2]

For these games, extant refinement criteria proposed for sequential games of incomplete information (see for example In-Koo Cho and David Kreps, 1987) do not directly apply. Furthermore, in the coordination games we consider, the equilibria will generally be proper (Roger Myerson, 1978) and perfect (Reinhard Selten, 1974). Nevertheless, a number of selection criteria have been proposed for games of this nature and these are reviewed in the next section. These criteria provide a number of hypotheses regarding equilibrium selection in coordination games.

The goal of this experiment is to address the following questions. First, is Nash equilibrium a good predictor of observed behavior in strategic or game settings? Second, when there are multiple Nash equilibria does

---

*Cooper, DeJong, and Forsythe: College of Business Administration, University of Iowa, Iowa City, IA 52242; Ross: Department of Economics, Carleton University, Ottawa, ON K1S 5B6, Canada. We thank Adil Abdalla for his help in conducting the experiments and Robert Pearson for his programming assistance. Helpful comments were received from seminar participants at the California Institute of Technology and the Universities of British Columbia, Illinois, Minnesota, California at Berkeley, California at Los Angeles and California at San Diego, Stanford University and Texas A&M University, as well as Raymond Battalio, Andrew Daughety, Paul Milgrom, Tom Palfrey, Richard Young and two anonymous referees. Financial support from the National Science Foundation, Price Waterhouse, University of Iowa, the Hoover Institution and the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

[1] This term has been used by Thomas Schelling (1960, ch. 4) to refer to games with multiple Nash equilibria yielding identical payoffs. He also used the term "collaboration games" for the same purpose. A standard example involves two automobiles approaching on a road. Each driver must select a side on which to drive. Equilibrium obtains when they select opposite sides. There are two such equilibria with identical payoffs. Here, we consider games in which the Nash equilibria are Pareto ranked, admitting a further need for coordination; namely, to select the "best" equilibrium.

[2] A slightly more extended discussion of these examples is included in Cooper, DeJong, Forsythe, and Ross (1989a).

one arise "naturally" as the outcome? That is, are there any particular characteristics of a game which tend to focus the players' attention on one of the equilibria?

The experimental design used to investigate these questions is discussed in Section I. The main results from our experiment are presented in Section II and can be summarized as follows. First, we find that the observed pattern of play is accurately predicted by the Nash equilibrium concept. Second, we find that the Pareto-dominant Nash equilibrium is not always the experimental outcome. This result is important in light of the argument often advanced that the Pareto-dominant outcome is a natural focal point. Instead, coordination failures can emerge in which the outcome is a Pareto-inferior Nash equilibrium.

We also provide evidence that dominated, cooperative strategies may influence the selection of an outcome from the set of Nash equilibria. Hence, the common practice of eliminating dominated strategies from a normal form game (discussed, for example, by Elon Kohlberg and Jean-Francois Mertens, 1986) may not be inconsequential. Section III provides an interpretation of our results.

I. Selection Criteria for Coordination Games

The critical element in coordination games is a "complementarity" between the strategy choice of a single agent and the strategy choice(s) of the other agent(s) (see Cooper and Andrew John, 1988). A specific game which illustrates a coordination problem is given in Figure 1. In this game, there are three strategies and two players with the payoff to the row player in each cell of the matrix. Since the game is symmetric, the payoff matrix for the column player is the transpose of this matrix. The pure strategy Nash equilibria are for the players to each choose strategy one, that is, (1,1), or for each to select strategy two, that is, (2,2). In these ordered pairs, the first element represents the strategy choice of the row player. These are both equilibria in that given the strategy chosen by the other player, neither player can profitably deviate. Note that at equilibrium (1,1), both players are worse off than at (2,2) though neither, acting independently, can do better. This is a coordination failure.

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	700
	2	250	550	1000
	3	0	0	600

FIGURE 1. ROW PLAYER'S PAYOFF MATRIX

If the players could cooperate and thus jointly determine their strategies, the best symmetric solution is for them to play strategy 3. Hereafter, we call strategy 3 the cooperative strategy and term (3,3) the cooperative outcome. Note also that strategy 3 is dominated by strategy 2 for the game in Figure 1.

The game in Figure 1 is representative of those in the experiment. It does not capture all of the economic detail in the examples of coordination failures. Rather, it reflects the essential qualitative properties of the coordination problem. We use this game to outline the alternative equilibrium selection criteria which have been proposed in the literature for games of this type.

We first allow for the possibility that the outcome is not a Nash equilibrium. For example, players may play the cooperative strategy or may randomly select strategies.

Hypothesis NE: The outcome will be a Nash equilibrium.

In the event that the outcome is in the set of Nash equilibria, one might argue that one of the equilibria is a focal point and hence represents a natural outcome for the game. That is, participants will believe that others will focus on one of the Nash equilibria, making it the outcome of their interaction. The issue of selection then reduces to a search for a "natural" focal point for the game.

Many game theorists (for example, Schelling, 1960 and John Harsanyi and

Selten, 1988) argue that the *Pareto-dominant* equilibrium is a natural focal point.[3] Under this selection hypothesis, coordination failures cannot occur as agents will naturally focus on the Pareto-dominant equilibrium, (2,2), and coordinate their actions to achieve that outcome.

Hypothesis PD: The Pareto-dominant Nash equilibrium will be selected.

We also consider the role of the dominated strategy in influencing the equilibrium selection process. While theorists have generally assumed that strictly dominated strategies can be excluded from the analysis of equilibrium selection (for example, Kohlberg-Mertens, 1986), experimental evidence on repeated prisoner's dilemma games suggests that players will sometimes play the dominated cooperative strategy, (for example, Robyn Dawes, 1980). This behavior may also have important implications for the selection of a Nash equilibrium. Further, to the extent that the selection of an equilibrium is influenced by payoffs associated with a dominated strategy, this provides evidence against any selection criteria which ignore these strategies.[4]

Hypothesis DSI: Dominated strategies are irrelevant to equilibrium selection.

II. The Experiment

In the experiment, players were asked to participate in a complete information, 3×3 bi-matrix game such as the one in Figure 1. Each player was paired with an anonymous opponent. One was designated the "row player" and the other the "column player." Each game was designed to be one of complete information because each player's payoff matrix was common knowledge and the numerical payoffs in each cell of the matrix represented a player's utility if the corresponding strategies were chosen.

We induced payoffs in terms of utility using the procedure introduced by Alvin Roth and Michael Malouf, (1979).[5] With this procedure, each player's payoff was given in points; these points determined the probability of the player's winning a monetary prize. At the end of each period of each game, we conducted a lottery where "winning" players received $1.00 and "losing" players received $0.00. The probability of winning, q, was given by dividing the points the player had earned by 1000. Thus, a player's expected utility at the outset of the game is given by:

$$(1)\quad U(\$1)q + U(\$0)(1-q) = q[U(\$1) - U(\$0)] + U(\$0).$$

Since expected utility is invariant with respect to linear transformations, q is a representation of the player's expected utility. Further, since q is linear in points, these points also denote a player's utility.

The experiment was conducted using seven cohorts of players, each consisting of eleven different players. All players were recruited from upper division undergraduate and MBA classes at the University of Iowa. Upon their arrival, players were seated at separate computer terminals and each was given a copy of the instructions for the experiment. These instructions are reproduced in Appendix A. Since these instructions were also read aloud, we assume that the information contained in them is common knowledge.

---

[3] Harsanyi and Selten (1988) uses the term "payoff dominance" to refer to the case in which payoffs are lower in one equilibrium than another for all players. Strictly speaking, Schelling [1960, Appendix C] argues that the Pareto-dominant equilibrium will be a focal point if there is only one combination of strategies yielding that outcome (as in Figure 1). Duncan Luce and Howard Raiffa (1957) go further, eliminating any Pareto-dominated strategy pair as "jointly inadmissible" and by defining a "solution" to be a Nash equilibrium involving only jointly admissible strategy pairs. Finally, Glenn Harrison and Jack Hirshleifer (1989), in an experimental study of public good provision, invoke Pareto dominance as a selection criterion in making their predictions.

[4] For example, Harsanyi and Selten (1988) consider *risk domination* in which the least risky equilibrium is thought to be a natural focal point. The riskiness of an equilibrium strategy is determined without considering strategies not used in any equilibrium. Thus variations in the payoff from the play of dominated strategies should, under this hypothesis, have no influence on the equilibrium that is selected.

[5] Both Roth and Malouf (1979) and Joyce Berg et al. (1986) have employed this procedure to control for risk preferences in experiments where other aspects of behavior were the primary focus of investigation. We have also conducted games similar to the games in this paper in which payoffs were in dollars rather than points. No apparent difference in behavior was noted.

Each player participated in a sequence of one-shot games against different anonymous opponents within his cohort. All pairing of players was done through the computer using a procedure described below. Since players reported their strategy choices through computer terminals, no player knew the identity of the player with whom he was currently paired, or the history of decisions made by any of the other players in the cohort.

Each cohort participated in two separate sessions.[6] In Session I, all players participated in ten symmetric one-shot dominant strategy games. The payoff matrix used in each of these games is shown in Game 1 in Figure 2. Since this game is symmetric, only the row player's payoff matrix is shown. During Session I, each player played one game against every other player. Since there was an odd number of players, one sat out in each period. Thus, Session I consisted of eleven periods. Also, players alternated being row and column players during the periods in which they were active participants. Game 1 was conducted for three reasons: first, to provide players with experience with experimental procedures, second, to see how well the dominant strategy equilibrium prediction performed, and finally, to inform players about the rationality of their opponents.

In Session II, all players participated in twenty additional one-shot games which differed from the game played in Session I. Each played against each other player twice: once as a row player and once as a column player. As in Session I, one player sat out in each period and players alternated between being row and column players during the periods in which they were participating.[7] Thus, Session II consisted of twenty-two periods.

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	320	440	500
	2	420	600	660
	3	180	360	420

Dominant Strategy Equilibrium at (2,2)

FIGURE 2. GAME 1 DOMINANT STRATEGY MATRIX

In both sessions, players were paired using a table which specified that in each eleven period sequence, each player would play every other player exactly once and would alternate between being the row and column player during the periods in which he was participating. The ordering of the players and consequently the pair assignments was randomized at the beginning of each session. This randomization procedure was adopted to prevent players from playing against the same opponent in the same order within each session. In the instructions players were told that they would play each other player once in Session I and twice in Session II (once as a row player and once as a column player).

Each cohort was given a different game to play in Session II. The payoff matrices for these games, called Games 2–8, are reported in Figures 3 and 4. Game 2 (see Figure 3) was asymmetric with a unique Nash equilibrium. This game was played to see if a unique Nash equilibrium, not supported by dominant strategies, is a good predictor of observed behavior. Therefore, the outcome for this game, as well as that from the dominant strategy game in Session I, provides evidence on Hypothesis NE.

---

[6] Each cohort completed the two sessions in about one hour. Payments to participants ranged from $5 to $20.

[7] We wanted the pairings to satisfy two conditions: (i) players were to alternate being row and column players and (ii) each player was to play each of the other players twice (in Session II). It is impossible to satisfy these two conditions with an even number of players. Having the player who sits out draw the lottery ticket may have also served the purpose of convincing players of the credibility of the lottery procedure.

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	525	555	585
	2	505	625	700
	3	385	550	625

Column Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	20	60	0
	2	110	420	495
	3	200	645	720

Nash Equilibrium at (2,3)

FIGURE 3. GAME 2 UNIQUE NASH EQUILIBRIUM MATRIX

Game 3

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	1000
	2	250	550	0
	3	0	0	600

Nash Equilibria at (1,1) and (2,2)

Game 4

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	700
	2	250	550	0
	3	0	0	600

Nash Equilibria at (1,1) and (2,2)

Game 5

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	700
	2	250	550	1000
	3	0	0	600

Nash Equilibria at (1,1) and (2,2)

Game 6

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	700
	2	250	550	650
	3	0	0	600

Nash Equilibria at (1,1) and (2,2)

Game 7

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	700
	2	250	550	0
	3	0	0	500

Nash Equilibria at (1,1) and (2,2)

Game 8

Row Player's Payoff Matrix

		Column Player		
		1	2	3
Row Player	1	350	350	1000
	2	250	550	0
	3	0	0	500

Nash Equilibria at (1,1) and (2,2)

FIGURE 4. MULTIPLE NASH EQUILIBRIUM GAMES

Games 3–8 (see Figure 4) are symmetric coordination games. These games were constructed to test the specific hypotheses given in the previous section. In coordination Games 3–8, the elements in each player's 2×2 principal minor are the same. Payoffs associated with a rival's play of strategy 3 vary from game to game, but strategy 3 is always dominated by strategy 1. The pure strategy Nash equilibria are at (1,1) and (2,2) and (2,2) dominates (1,1) in all cases. In Games 3–6, the cooperative outcome occurs at (3,3) and in Games 7 and 8, it occurs at (2,2).

With this design in mind, we can outline our tests of the hypotheses presented in the previous section. Hypothesis NE will not be rejected if we observe Nash equilibria in each of our games. Should we fail to reject this hypothesis, we will test the alternative selection criteria. Hypothesis PD will be rejected if (1,1) is observed. Evidence that dominated strategies matter in the selection of an equilibrium will serve to refute Hypothesis DSI and, consequently, selection criteria which require that outcomes be independent of payoffs from dominated strategies.

III. Results

We present our results in several parts. First, we examine the behavior in games with unique equilibria (Games 1 and 2). Second, we analyze the data from the coordination games, Games 3–6, and examine the predictive ability of the selection criteria presented in Section II. Games 7 and 8 are discussed in the following section.[8]

A. *Games 1–2*

The data from Games 1 and 2 are presented in the upper and lower panels of Figure 5. Since Game 1 is a symmetric game, we tested to see if row players chose different strategies than column players. Using Fisher's exact probability test for player differences,[9] we found no effects approaching any reasonable level of significance; therefore, we pooled the data across row and column players. Since Game 2 is an asymmetric game, we attempted no such pooling. The row players' choices in this game are presented on the right-hand side of panel b and the column players' choices are shown on the left.[10] We next tested to see if the observed pattern of play was independent from one period to the next. Again using a Fisher's exact test, we were unable to reject the null hypothesis of independence at any reasonable level of significance. Thus, the remaining analysis pools the results within each of these two games.

From Figure 5 it is apparent that there is strong support for the dominant strategy prediction in Game 1 and the Nash equilibrium prediction in Game 2. From the outset, players almost always chose their Nash equilibrium strategies in these two games.

B. *Games 3–6*

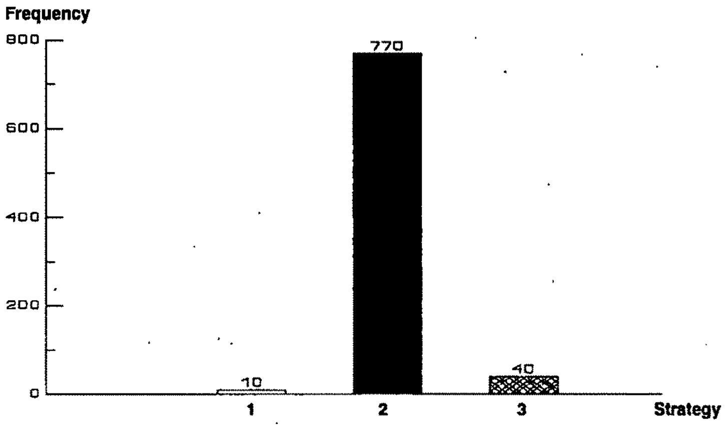
We next analyze the first four coordination games. Recall that these are all symmetric games. As with Game 1, we tested whether row players behaved differently than column players. Since we again found no significant differences, we pooled the data across both types of players. The data are not independent across periods, however. Using Fisher's exact test, we can reject the hypothesis that the data from the first 11 periods is the same as the data from the last 11 periods at the 0.05 level. Using only the last half of the

---

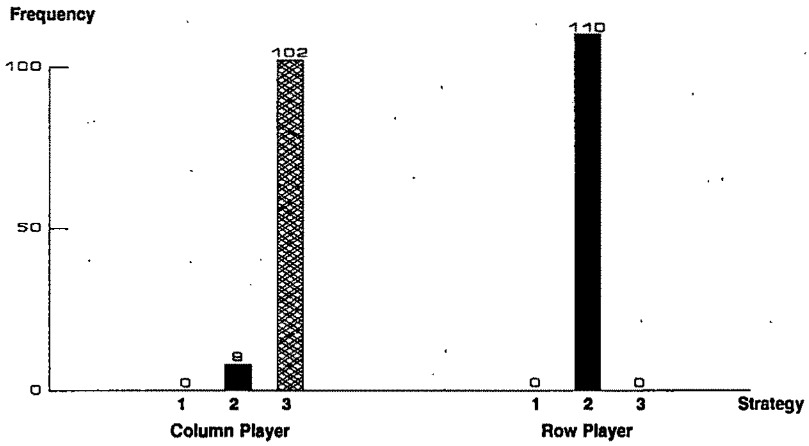
[8] The complete history of play for each game is available from the authors upon request.

[9] We use Fisher's exact test throughout this paper to test for the statistical significance of differences (see Kendall and Stuart, Vol. II (1979) 584–86). For example, in Game 3, the strategies 1, 2, and 3 were chosen 157, 44, and 19 times, respectively. The row players choose these strategies 79, 33, and 8 time, respectively. We ask how likely we are to see these choices if they are independent of whether the player was designated a row or column player. The exact probability of observing this data is 0.737.

[10] Throughout the discussion of results, we describe individual play rather than equilibrium outcomes as strategy choices are independent of rivals' identities given the information structure.



a. Game 1

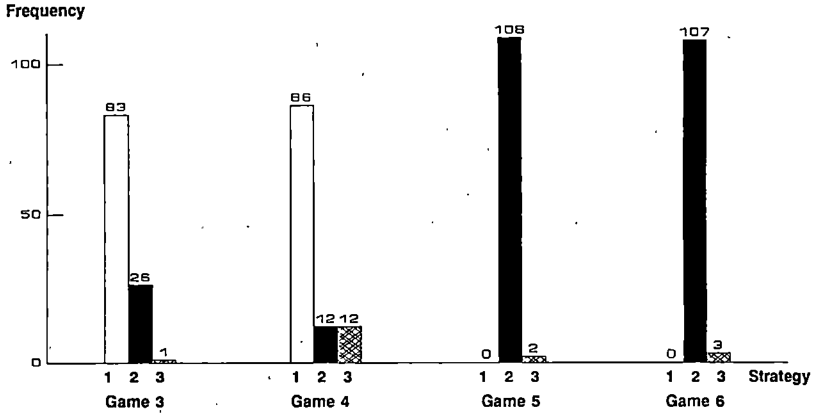


b. Game 2

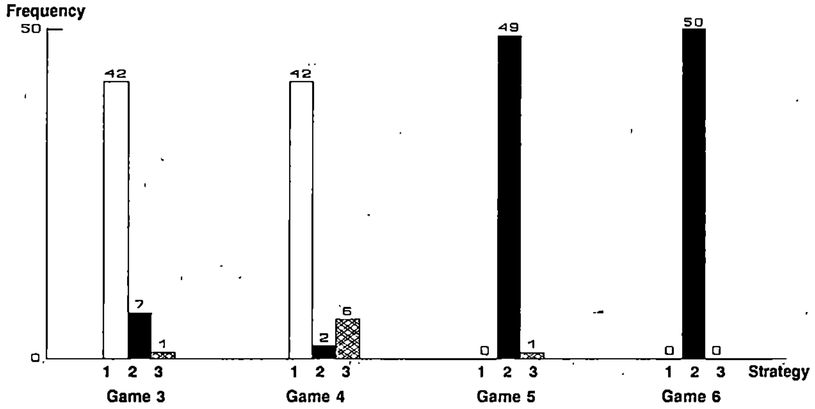
FIGURE 5a. GAME 1, b. GAME 2

data from each game, this time independence hypothesis cannot be rejected at a significance level of 0.05. However, since Game 4 barely passed this test, we examined the data from the last five periods of each game. In this latter case, we find that the time independence hypothesis cannot be rejected at any significance level up to 0.40. Thus, we report our data in two ways: pooled over the last eleven periods and pooled over the last five periods.

These pooled data are summarized in Figure 6. From this figure it can be seen that players overwhelmingly chose a single strategy in each of these games: strategy 1 in Games 3 and 4 and strategy 2 in Games 5



a. Last 11 Periods



b. Last 5 Periods

FIGURE 6a. LAST 11 PERIODS, b. LAST 5 PERIODS

nd 6. These data are consistent with the ypothesis that the outcomes will be a Nash quilibrium.[11,12]

We can use the data to refute several additional hypotheses. Consider first Hypothesis PD and recall that (2,2) is the Pareto-dominant Nash equilibrium strategy in these games. The data from Games 3 and 4 can be used to reject the hypothesis. Further, the data support the hypothesis that

---

[11] A more formal test of the hypotheses presented in he previous section can be found in Cooper, DeJong, orsythe, and Ross (1989a). For each of the respective ames, we were unable to reject the hypothesis that trategy 1 was chosen at least 95 percent of the time in ames 3 and 4 and strategy 2 was chosen at least 95 ercent of the time in Games 5 and 6 in the last five eriods of play.

[12] We focus here on pure strategy Nash equilibria. In ll of Games 3–8 there is a unique mixed strategy equilibrium in which each player plays strategy one with probability 2/3 and strategy two with probability 1/3. The data presented in Figure 6 reveal that we can reject this as the outcome over the last 11 and last 5 periods in all of these games.

(1,1) is the experimental outcome. Thus, Games 3 and 4 provide examples of coordination failures.[13]

Hypothesis DSI can be tested by comparing the outcomes from these games. According to this hypothesis, the experimental outcomes should be the same in Games 3 to 6, since they differ only in the payoffs a player receives when his opponent plays a dominated strategy. However, the Pareto-inferior Nash equilibrium outcome was observed in Games 3 and 4, while the Pareto-superior Nash equilibrium outcome was observed in Games 5 and 6.[14] Thus, comparing these outcomes leads us to reject Hypothesis DSI: dominated strategies can influence the selection of an outcome from the set of Nash equilibria.

IV. Interpretation

These results provide evidence against Pareto-dominance as a selection criteria for coordination games and indicate that the cooperative, dominated strategy influences equilibrium selection. That is, by varying the payoffs associated with a rival's play of strategy 3, we can induce variations in the selection of an equilibrium. At the very least, the evidence implies that at some point during the play of these games, some participants placed positive probability on an opponent playing the cooperative, dominated strategy. If this was not the case, then variations in the payoff associated with an opponent playing this strategy could not have affected the outcome. From these observations, two questions arise. First, why is strategy 3 so important in these games? Second, how did the play evolve toward the equilibrium that was selected?

We investigated whether strategy 3 is important because it is the cooperative strategy or whether players place prior probability weight on a dominated strategy even when it does not support the cooperative outcome. That is, a player might believe that his opponent will be "cooperative" or will be "irrational" where the latter term refers to the play of a dominated strategy by a self-interested player.[15]

To separate the two explanations for our observations in Games 3–6, we designed Games 7 and 8 (see Figure 4) in which the cooperative outcome was not supported by dominated strategies. Game 7 (8) was constructed from Game 4 (3) by reducing each player's payoff from playing strategy 3 when his opponent plays strategy 3 from 600 to 500. For each of these games, this moves the cooperative outcome from (3,3) to (2,2) which is also a Nash equilibrium. If players place the same prior probability on their opponent playing the dominated strategy in Game 7 (8) as they did in Game 4 (3), the outcome observed in these two games should be the same. On the other hand, if players place positive probability on their opponent playing the cooperative strategy rather than the dominated strategy, players should expect that their opponent is more likely to play (2,2) in Game 7 (8) than in Game 4 (3). This should increase the likelihood that (2,2) is observed in Games 7 and 8.

The pooled data for Games 7 and 8 are summarized in Figure 7. The Pareto-superior Nash equilibrium, (2,2), was observed in

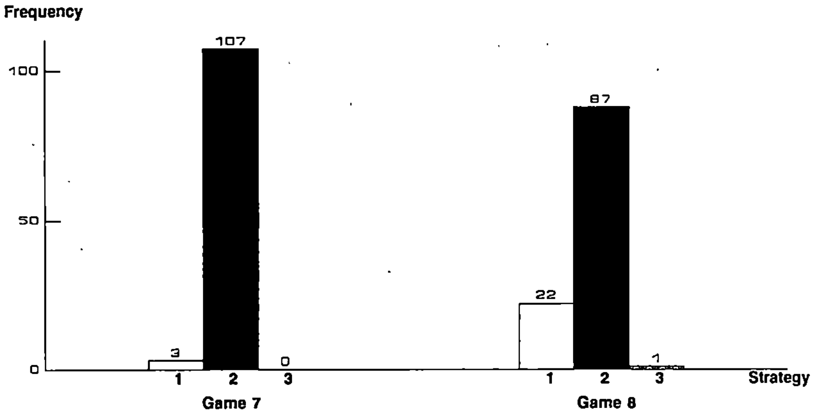
---

[13] Since running these initial treatments, we have replicated the results for Game 3 two additional times. In an experiment based on Bryant's (1983) model of coordination failures, John Van Huyck et al. (1987) also find that the Pareto-inferior Nash equilibrium may be selected.

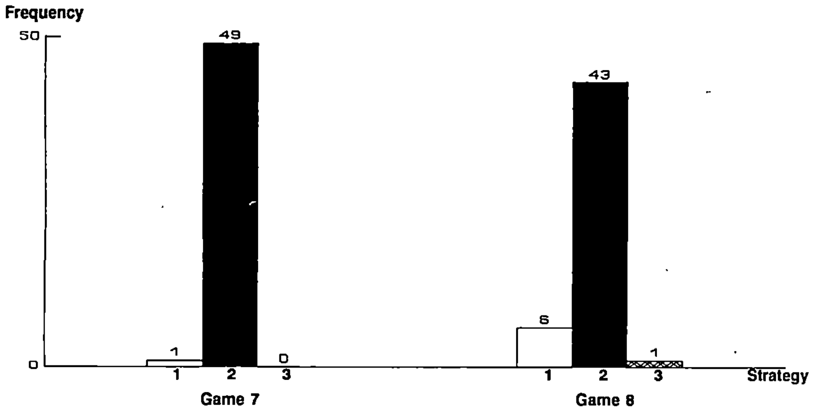
However, we want to indicate a word of caution about the saliency of the $1 lottery prize. Since conducting the initial treatment and replications (about two years ago), we've obtained identical results in four replications using a $2 prize. With a $1 price, the results are qualitatively similar but not identical.

[14] The Pareto-superior outcome was also obtained in an additional game we conducted using the 2×2 principal minor of Games 3–6. For the 22 periods, players selected strategy 2 108 out of 110 opportunities. When compared with Games 3–4, this game provides additional evidence regarding the influence of a dominated strategy on equilibrium selection.

[15] This should be distinguished from the play of a cooperative, dominated strategy by a player whose utility depends, in part, on the payoffs of others.



a. Last 11 Periods



b. Last 5 Periods

FIGURE 7a: LAST 11 PERIODS, b. LAST 5 PERIODS

these games.[16] Thus, when strategy 3 no longer supports the cooperative outcome, players evidently place a lower probability on its play by their rival. This is verified by the observed play: in all 22 periods of Games 3–6 the cooperative, dominated strategy is chosen 11 percent of the time while in Games 7–8, the dominated strategy is played only 1.8 percent of the time.

To better understand the observed differences in play in Games 3–6, Figure 8 presents the data for all 22 periods of play in each game. The play of strategy 3 occurs more frequently in early periods of all four games. As indicated by the figure, this play tends to disappear in all games except Game

---

[16] There are some differences in play between these two games. In particular, strategy 1 was played 22 times in Game 8 but only 3 times in Game 7 in the last 11 periods.

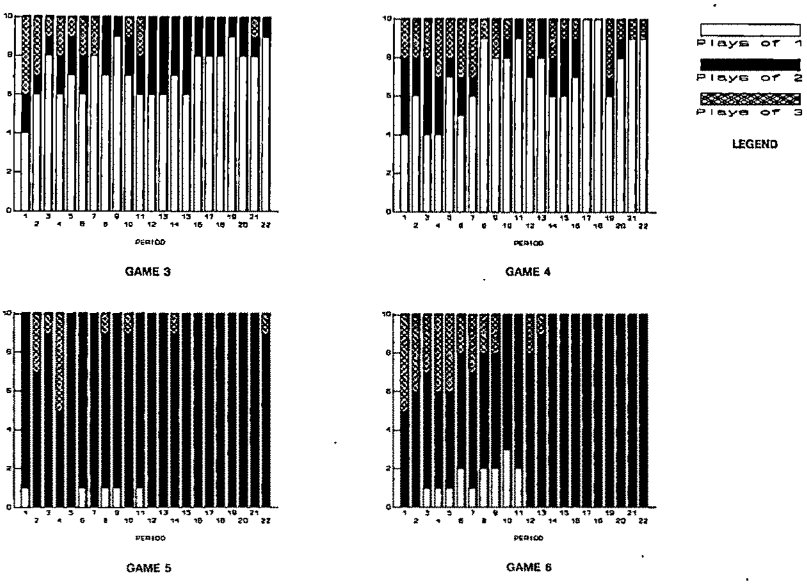


FIGURE 8. GAMES 3, 4, 5, AND 6

TABLE 1—FREQUENCY OF STRATEGY CHOICES BY OPPONENT'S PLAY IN PREVIOUS PERIOD: Periods 2–11, Games 3–6

		Opponent's Play in Previous Period			
Game	Strategy	1	2	3	Total
	1	48	9	12	69
3	2	11	4	1	16
	3	8	3	3	14
	Total	67	16	16	99
	1	44	8	13	65
4	2	8	5	2	15
	3	8	5	6	19
	Total	60	18	21	99
	1	0	4	0	4
5	2	4	72	9	85
	3	0	8	2	10
	Total	4	84	11	99
	1	2	9	4	15
6	2	8	35	17	60
	3	3	13	8	24
	Total	13	57	29	99

4 in which strategy 3 is played a number of times in later periods. There is little initial play of strategy 2 (1) in Games 3 and 4 (5 and 6). In Games 3 and 4, the play of strategy 1 increases over time. In Games 5 and 6, the play of strategy 2 increases over time and strategy 1 is never observed in the last 11 periods of play.

Table 1 summarizes individual play in the first 11 periods conditional on opponents' play in the previous period.[17] The point of this table is to illustrate the evolution of play to strategy 1 (2) in Games 3 and 4 (5 and 6). For Games 3 and 4, an opponent's play of strategy 3 appears to lead to the play of strategy 1 in the next period. Similarly, in Games 5 and 6, an opponent's play of 3 leads to the play of strategy 2 in the next period.

A model consistent with these observations must explain the observed play of the cooperative strategy and the selection process. One model assumes the presence of some altruistic players.[18] Suppose, for example, that altruists want to maximize the minimum element of the payoff pair so that strategy 3 becomes a best response to a rival's play of strategy 3, and outcome (3,3) represents an equilibrium for two such players.[19] Equilibrium selection would then depend upon the actual proportion of players who are altruistic as well as the priors players hold about this proportion. For example: self-interested players who expect to meet a sufficient number of altruists playing strategy 3, will respond by playing strategy 1 in Games 3, 4, and 6, and strategy 2 in Game 5. Altruists who find the probability of meeting another altruist playing strategy 3 too low, will switch to whatever the self-interested players are playing, yielding an equilibrium at (1,1) or (2,2). If we use the actual proportion of play observed in the first 11 periods of each game as proxies for the actual beliefs held by players, this model predicts the equilibria observed in Games 3–8.

V. Conclusions

Our principal concern in this research was to answer the following question: What happens when players play a noncooperative game with multiple Nash equilibria? We further refined the question by focusing on games with Pareto-rankable equilibria called coordination games. Coordination games are of significant economic interest; they describe the strategic interactions present in models of double moral hazard and networks in the industrial organization field and coordination failures in macroeconomics. In terms of the outcomes in these games, several possibilities suggest themselves. For example, we could observe the Pareto-dominant equilibrium as the outcome, randomized strategies, chaos, or even cooperation.

Our experimental evidence strongly supports the hypothesis that the outcome will be from the set of Nash equilibria. It does not support, however, any of the other proposed hypotheses regarding equilibrium selection. First, the Pareto-dominant equilibrium was not always selected in the games. Second, we found evidence that variations in payoffs from a rival's play of a dominated strategy can influence equilibrium selection. As a result, we can reject the hypotheses based upon Pareto dominance and the irrelevance of dominated strategies; variations in a player's payoff from an opponent's play of a cooperative, dominated strategy influences equilibrium selection.

The importance of cooperative play has also been identified in the experimental literature on finitely repeated prisoner's dilemma games.[20] There are a number of theoretical models which can explain this cooperative play including that discussed by Kreps et al.

---

[17] This allows us to look at the most naive type of dynamics: best responding to an opponent's play in the previous period.

[18] Alternatively, Kreps et al. explain observed cooperation in finitely repeated prisoner's dilemma games by modeling it as a game of incomplete information. If players believe that there is a small chance their opponent will adopt a tit-for-tat strategy, it may be optimal for them to respond for some time, with a similar strategy, leading to cooperation. The strategies which support this outcome are not feasible in our environment.

[19] This is only one of many ways to model altruistic behavior. Thomas Palfrey and Howard Rosenthal (1987) introduce a different model of altruism in their work on experimental public goods contribution games. In their model, altruists enjoy extra utility from being "cooperative" (in their model by contributing more toward the provision of the public good) which is independent of the choices made by the other players.

[20] See, for example, Dawes (1980), Dawes and John Orbell (1982) and Anatol Rapoport and Albert Chammah (1965).

(1982) and models with altruistic players. We are unaware of any evidence which distinguishes between these competing theories of cooperative play.

A second line of research is to consider the robustness of observed coordination failures to alternative institutions. Joseph Farrell (1987) suggests that nonbinding preplay communication can overcome coordination problems while Kohlberg-Mertens (1986) suggest that allowing one player to choose between playing a coordination game and receiving a certain outcome can lead to play of the Pareto-dominant Nash equilibrium.[21]

APPENDIX A: INSTRUCTIONS

General

You are about to participate in an experiment in the economics of decision making. If you follow these instructions carefully and make good decisions, you might earn a considerable amount of money which will be paid to you in cash at the end of the experiment.

The experiment will consist of a series of separate decision making periods. Each period consists of two phases. In Phase I you will be paired with another person and, based upon your combined actions, you will be able to earn *points*. In Phase II, you will have the opportunity to earn dollars based upon the points you earn in Phase I. We begin by describing Phase II so that you understand how the points you earn affect the number of dollars you earn. Then, we describe Phase I in detail so that you understand how to earn points.

Phase II *Instructions*

At the end of Phase I, you will have earned between 0 and 1000 points according to the rules we will discuss below. The number of dollars you earn in Phase II will depend partly on the number of points you earned in Phase I and partly on chance. Specifically, we have a box which contains lottery tickets numbered 1 to 1000. In Phase II, a ticket will be randomly drawn from the box. If the number on the ticket IS LESS THAN OR EQUAL TO the number of points you have earned in Phase I, you WIN $1.00. If the number on this ticket IS GREATER THAN the number of points you have earned in Phase I, you WIN $0.00. For example, if you have 600 points, you will have a 60 percent chance of winning $1.00. Notice that the more points you have, the larger will be your chance of winning the $1.00 prize.

Phase I *Instructions*

In each decision-making period, you will be paired with another person. One of you will be designated player *B* and the other will be designated player *S*. At the beginning of the period, both player *B* and player *S* must separately and independently select an action. The combined actions of player *B* and player *S* jointly determine the number of points earned by player *B* and the number of points earned by player *S*.

You will alternate from being player *B* to being player *S* from one period to the next. Since there is not an even number of people participating in this experiment, you will occasionally be required to not participate during a particular period. When this is the case, you will receive a message on your terminal which states:

"FOR PERIOD ____, YOU ARE SITTING OUT."

In the periods in which you are participating you will receive a message stating:

"FOR PERIOD ____, YOU ARE A *B* PLAYER."

or

"FOR PERIOD ____, YOU ARE AN *S* PLAYER."

You will be participating in a series of separate sessions during today's experiment. During the current session, you will play against each person once—as either player *B* or player *S*. However, you will not know the identification of the person you are playing against in any period. Similarly, nobody in your decision-making pair will know your identification in any period. Further, you will not be told who these people are either during or after the experiment.

In your folder you will find a set of record sheets. On these sheets you will indicate, based on the message previously received on your terminal, whether you are player *B*, player *S*, or not participating for each period. The points that you earn in each period will be determined by the rules given below.

Specific Instructions for Player B

In this part of the instructions we will be referring to specific number of points. These numbers are the same as you will be using in the first session of today's experiment.

In those periods in which you are player *B*, you and player *S* must separately and independently decide on actions which will jointly determine the number of points earned by you and the number of points earned by player *S*. As player *B*, you may either choose action *B*1, action *B*2, or action *B*3. Similarly, player *S* may choose action *S*1, action *S*2, or action *S*3. The number of points earned by you is given by the following table for each pair of actions you and player *S* might select:

---

[21] Cooper, DeJong, Forsythe, and Ross (1989b) provide some preliminary results indicating that allowing communication in Game 3 may not lead to play of the Pareto-dominant Nash equilibrium. Van Huyck et al. (1989) present results on coordination games with pre-play auctions which support the predictions of Kohlberg-Mertens.

NUMBER OF POINTS EARNED BY PLAYER *B*

		S's Actions		
		*S*1	*S*2	*S*3
B's Action	*B*1	320	440	500
	*B*2	420	600	660
	*B*3	180	360	420

To read this table, suppose that you chose action *B*2 and player *S* chose action *S*1. You would then earn 420 points. Similarly, suppose that you chose action *B*1 and player *S* chose action *S*3. You would then earn 500 points. In a like manner, you can use this table to determine the number of points you would earn for all other pairs of actions you and player *S* may select. *S* players also earn points depending upon the type of action they select. These are given in the next section of the instructions.

When you select an action, enter the action chosen into the computer via your terminal and record the action chosen on your record sheet. Once both you and player *S* have selected your actions and entered them into the computer via your terminals, the computer will determine the number of points earned by you based on the table given above. The result is then sent to you via your terminal. The message will look like the one below:

PERIOD POINTS ARE ____.

At the end of the period, you are to record your point earnings for Phase I on your record sheet. Make sure you check your earnings in points against the computer's calculations. The computer will also inform you about the action taken by player *S*. Make sure you record this information on your record sheet.

Specific Instructions to Player S

In those periods in which you are player *S*, you and player *B* must separately and independently decide on actions which will jointly determine the number of points earned by you and the number of points earned by player *B*. As player *S*, you may either choose action *S*1, action *S*2, or action *S*3. The number of points earned by you is given by the following table for each pair of actions you and player *B* might select:

NUMBER OF POINTS EARNED BY PLAYER *S*

		S's Action		
		*S*1	*S*2	*S*3
B's Action	*B*1	320	420	180
	*B*2	440	600	360
	*B*3	500	660	420

To read this table, suppose that player *B* chose action *B*2 and you chose action *S*1. You would then earn 440 points. Similarly, suppose that player *B* chose action *B*1 and you chose action *S*3. You would then earn 180 points.

When you select an action, enter the action chosen into the computer via your terminal and record the action chosen on your record sheet. Once both you and player *B* have selected your actions and entered them into the computer via your terminals, the computer will determine the number of points earned by you based on the table given above. The result is sent to you via your terminal. The message will look like the one below:

PERIODS POINTS ARE ____.

At the end of the period, you are to record your point earnings for Phase I on your record sheet. Make sure you check your earnings in points against the computer's calculations. The computer will also inform you of the action taken by player *B*. Make sure you record this information on your record sheet.

Phase II *Recording Rules*

After completing your Phase I record sheet for a given decision-making period, you are to use your profit sheet to record the dollars you earn in Phase II. First, record your Phase I point earnings in the row corresponding to the number of the period that is currently being conducted. The person who sat out in this period will then be asked to draw a lottery ticket from the box. Before he/she returns the ticket to the box, the number on the ticket will be announced. You should record the number of the ticket in the second column of your profit sheet. If the number drawn IS LESS THAN OR EQUAL TO the number of points earned in Phase I, circle $1.00 in the next column; otherwise circle $0.00 in that column. Pay careful attention to what you circle. Any erasure will invalidate your earnings for the period. If you do make a mistake and circle the wrong number, call it to the experimenter's attention.

At the end of the session, add up your total profit in dollars and record this sum in row 23 of your profit sheet. All dollars on hand at the end of the session in excess of $2.00 dollars are yours to keep. Subtract this number, which is on row 24, from your total dollars on row 23 and record this difference on row 25. This is the amount of dollars you have earned in this session.

In summary, your earnings in the experiment will be the total of the amounts you win in all Phase II lotteries. The amount of money you earn will depend partly upon luck and partly upon whether you have made good decisions in Phase I. Notice that the more points you earn in Phase I, the more likely you will win in Phase II. Are there any questions?

Instructions for Session II

This session of the experiment will again consist of a series of separate decision-making periods. The session will be conducted in exactly the same way as the previous session except you will play against each person twice, once as player *S* and once as player *B*.

All dollars you earn in this session in excess of $0.00 dollars are yours to keep.

SAMPLE INFORMATION SHEET

Player B

Number of Points Earned by Player B

		S's Action		
		*S*1	*S*2	*S*3
B's Action	*B*1	320	440	500
	*B*2	420	600	660
	*B*3	180	360	420

Player S

Number of Points Earned by Player *S*

		S's Action		
		*S*1	*S*2	*S*3
B's Action	*B*1	320	420	180
	*B*2	440	600	360
	*B*3	500	660	420

REFERENCES

Berg, Joyce E. et al., "Controlling Preferences for Lotteries on Units of Experimental Exchange," *Quarterly Journal of Economics*, May 1986, *101*, 281–306.

Bryant, John, "A Simple Rational Expectations Keynes-Type Model," *Quarterly Journal of Economics*, August 1983, *98*, 525–59.

Cho, In-Koo and Kreps, David, "Signalling Games and Stable Equilibria," *Quarterly Journal of Economics*, May 1987, *102*, 179–222.

Conover, W. J., *Practical Nonparametric Statistics*, New York: Wiley & Sons, 1971.

Cooper, Russell et al., (1989a) "Selection Criteria in Coordination Games: Some Experimental Results," University of Iowa Working Paper, 1989.

______, (1989b) "Communication in Coordination Games," University of Iowa, mimeo., 1989.

Cooper, Russell and John, Andrew, "Coordinating Coordination Failures in Keynesian Models," *Quarterly Journal of Economics*, August 1988, *103*, 441–63.

______ **and Ross, Thomas, W.,** "Product Warranties and Double Moral Hazard," *Rand Journal of Economics*, Spring 1985, *16*, 103–113.

Dawes, Robyn M., "Social Dilemmas," *Annual Review of Psychology*, 1980, *31*, 169–93.

______ **and Orbell, John M.,** "Cooperation in Social Dilemma Situations: Thinking About It Doesn't Help," in *Research in Experimental Economics, Volume II*, Vernon L. Smith, ed., Greenwich: JAI Press, 1982, 167–73.

Diamond, Peter, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, *90*, 881–94.

Farrell, Joseph, "Cheap Talk, Coordination, and Entry," *Rand Journal of Economics*, Spring 1987, *18*, 34–39.

Harrison, Glenn W. and Hirshleifer, Jack, "An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods," *Journal of Political Economy*, February 1989, *97*, 201–225.

Harsanyi, John C. and Selten, Reinhard, *A General Theory of Equilibrium Selection in Games*, Cambridge: MIT Press, 1988.

Heller, Walter, "Coordination Failure under Complete Markets with Applications to Effective Demand," in *Equilibrium Analysis, Essays in Honor of Kenneth J. Arrow, Volume II*, Walter Heller, Ross Starr, and David Starrett, eds., Cambridge: Cambridge University Press, 1986.

Katz, Michael and Shapiro, Carl, "Network Externalities, Competition and Compatibility," *American Economic Review*, June 1985, *75*, 424–40.

Kendall, Maurice and Stuart, Alan, *The Advanced Theory of Statistics, Vol. 2*, New York: Macmillan, 1979.

Kiyotaki, Nobuhiro, "Implications of Multiple Expectations Equilibria Under Monopolistic Competition," *Quarterly Journal of Economics*, November 1988, *103*, 695–714.

Kohlberg, Elon and Mertens, Jean-Francois, "On the Strategic Stability of Equilibria," *Econometrica*, September 1986, *54*, 1003–38.

Kreps, David et al., "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory*, August 1982, *17*, 245–52.

Luce, R. Duncan and Raiffa, Howard, *Games and Decisions*, New York: Wiley & Sons, 1957.

Myerson, Roger, "Refinements of the Nash Equilibrium Concept," *International Journal of Game Theory*, 1978, *7*, 73–80.

Palfrey, Thomas and Rosenthal, Howard, "Private Incentives and Social Dilemmas: The Effects of Incomplete Information and Altruism," mimeo., 1987; forthcoming, *Journal of Public Economics*.

Rapoport, Anatol and Chammah, Albert M., *Prisoner's Dilemma*, Ann Arbor: University of Michigan Press, 1965.

Roth, Alvin E. and Malouf, Michael W. K., "Game-Theoretic Models and the Role of Bargaining," *Psychological Review*, 1979, *86*, 574–94.

Schelling, Thomas, *The Strategy of Conflict*, Cambridge: Harvard University Press, 1960.

_____, *Micromotives and Macrobehavior*, New York: Norton & Company, 1978.

Selten, Reinhard, "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1974, *4*, 25–55.

Van Huyck, John B. et al., "Coordination Failure, Strategic Uncertainty, and Convergent Expectations," mimeo., Department of Economics, Texas A&M University, July 1987.

_____, "Asset Markets as an Equilibrium Selection Mechanism: Coordination Failure, Game Form Auctions, and Forward Induction," Texas A&M University Working Paper No. 88-40, February 1989.

Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure

By JOHN B. VAN HUYCK, RAYMOND C. BATTALIO, AND RICHARD O. BEIL*

Deductive equilibrium methods—such as Rational Expectations or Bayesian Nash Equilibrium—are powerful tools for analyzing economies that exhibit strategic interdependence. Typically, deductive equilibrium analysis does not explain the process by which decision makers acquire equilibrium beliefs. The presumption is that actual economies have achieved a steady state. In economies with stable and unique equilibrium points, the influence of inconsistent beliefs and, hence, actions would disappear over time, see Robert Lucas (1987). The power of the equilibrium method derives from its ability to abstract from the complicated dynamic process that induces equilibrium and to abstract from the historical accident that initiated the process.

Unfortunately, deductive equilibrium analysis often fails to determine a unique equilibrium solution in many economies and, hence, often fails to prescribe or predict rational behavior. In economies with multiple equilibria, the rational decision maker formulating beliefs using deductive equilibrium concepts is uncertain which equilibrium strategy other decision makers will use and, when the equilibria are not interchangeable, this uncertainty will influence the rational decision-maker's behavior. Strategic uncertainty arises even in situations where objectives, feasible strategies, institutions, and equilibrium conventions are completely specified and are common knowledge. While multiple equilibria are common in theoretical analysis, consideration of specific economies suggests that many equilibrium points are implausible and unlikely to be observed in actual economies.

One response to multiple equilibria is to argue that some Nash equilibrium points are not self-enforcing and, hence, are implausible, because they fail to satisfy one or more of the following refinements: elimination of individually unreasonable actions, sequential rationality, and stability against perturbations of the game—see Elon Kohlburg and Jean-Francois Mertens (1986) for examples and references. Equilibrium refinements determine when an outcome that is *already expected* would be implemented by rational decision makers.

In general, many outcomes will satisfy the conditions of a given equilibrium refinement. The equilibrium selection literature attempts to determine which, if any, self-enforcing equilibrium point will be expected. A satisfactory theory of interdependent decisions must not only identify the outcomes that are self-enforcing when expected but also must identify the expected outcomes. Consequently, a theory of equilibrium selection would be a useful complement to the theory of equilibrium points.

The experimental method provides a tractable and constructive approach to the equilibrium selection problem. This paper studies a class of tacit pure coordination games with multiple equilibria, which are strictly Pareto ranked, and it reports experiments that provide evidence on how human subjects make decisions under conditions of strategic uncertainty.

---

*Department of Economics, Texas A&M University, College Station, TX 77843; Department of Economics, Texas A&M University, College Station, TX 77843; and Department of Economics, Auburn University, Auburn, AL 36849, respectively. Mike Baye, John Bryant, Colin Camerer, Russell Cooper, Vincent Crawford, John Haltiwanger, Pat Kehoe, Tom Saving, Steve Wiggins, Casper de Vries, a referee, and seminar participants at the NBER/FMME 1987 Summer Institute, the 1988 meetings of the Economic Science Association, the University of New Mexico, and Texas A&M University made constructive comments on earlier versions of this paper. Sophon Khanti-Akom, Kirsten Madsen, and Andreas Ortmann provided research assistance. The National Science Foundation, under grants no. SES-8420240 and SES-8911032, the Texas A&M University Center for Mineral and Energy Research, and the Lynde and Harry Bradley Foundation provided financial support. A Texas A&M University TEES Fellowship has supported Battalio.

I. A Pure Coordination Game

To focus the analysis consider the following tacit coordination game, which is a strategic form representation of John Bryant's (1983) Keynesian coordination game. The baseline game is defined as follows: Let $e_1,\dots,e_n$ denote the actions taken by n players. The period game A is defined by the following payoff function and strategy space for each of n players:

$$\pi(e_i,\underline{e}_i)=a[\min(e_i,\underline{e}_i)]-be_i, \quad a>b>0, \tag{1}$$

where $\underline{e}_i$ equals $\min(e_1,\dots,e_{i-1},e_{i+1},\dots,e_n)$. Actions are restricted to the set of integers from 1 to $\bar{e}$. The players have complete information about the payoff function and strategy space and know that the payoff function and strategy space are common knowledge.[1]

If the players could explicitly coordinate their actions, the—real or imagined—planner's decision problem would be trivial. Given $a-b$ greater than 0, each player should choose the maximum feasible action, $\bar{e}$. Moreover, a negotiated "pregame" agreement to choose $\bar{e}$ would be self-enforcing. Unlike games with incentive problems, here the first best outcome is an equilibrium point. However, when the players cannot engage in "pregame" negotiation they face a nontrivial coordination problem.

Suppose that the players attempt to use the Nash equilibrium concept to inform their strategic behavior in the tacit coordination game A. A player's best response to $\underline{e}_i$ is to choose e_i equal to $\underline{e}_i$. By symmetry it follows that any n-tuple $(e,\dots,e)$ with $e\in\{1,2,\dots,\bar{e}\}$ satisfies the mutual best response property of a Nash equilibrium point. All feasible actions are potential Nash equilibrium outcomes. The Nash concept neither prescribes nor predicts the outcome of this tacit coordination game.

(Standard equilibrium refinements do not reduce the set of equilibria. For example, the equilibria are strict—each player has a unique best response—and, hence, trembling-hand perfect.)

II. Coordination Problems and Equilibrium Selection Principles

The analysis in Section I follows convention and abstracts from the equilibrium selection problem. However, a rational player using deductive equilibrium concepts confronts two nontrivial coordination problems in period game A. First, players may fail to correctly forecast the minimum, $\underline{e}_i$, and, hence, regret their individual choice, that is, $e_i\neq\underline{e}_i$. This type of coordination failure results in disequilibrium: outcomes that do not satisfy the mutual best-response property of an equilibrium.

An equilibrium selection principle identifies a subset of equilibrium points according to some distinctive characteristic. An interesting conjecture is that decision makers use some selection principle to identify a specific equilibrium point in situations involving multiple equilibria. This selection principle would solve the problem of coordinating on a specific equilibrium point. Hence, the outcome will satisfy the mutual best-response property of an equilibrium.

A second coordination problem arises when the equilibria can be Pareto ranked. In such situations, all players may give a best response, but, nevertheless, implement a

---

[1] Apparently, this game is similar to Rousseau's "stag hunt" parable, which he used to motivate his analysis of the social contract, see Crawford (1989, p. 4). In the stag hunt game, each hunter in a group must allocate effort between hunting a stag with the group and hunting rabbits by himself. Let e_i denote effort expended on the stag hunt. Since stag hunting in that era required the coordinated effort of *all* the hunters, the probability of successfully hunting a stag depends on the smallest e_i. The parameter a in equation (1) reflects the benefits of participating in the stag hunt: eating well should the hunt succeed. Hunting rabbits does not require coordination with the other hunters. The parameter b in equation (1) reflects the opportunity cost of effort allocated to the stag hunt that could have been allocated to rabbit hunting: a meal—however, meager.

Pareto dominated equilibrium, that is, $\min(e_1,\dots,e_n) \neq \bar{e}$. While not regretting their individual choice, they regret the equilibrium implemented by these individual choices. Consequently, the outcome results in coordination failure. What equilibrium selection principles could a player use to resolve these two coordination problems?[2]

Deductive selection principles select equilibrium points based on the description of the game. Deductive selection principles preserve the equilibrium method's desirable property of independence from historical accidents and from complicated dynamic processes. Inductive selection principles select equilibrium points based on the history of some pregame.[3] Hence, inductive selection principles are not independent of accident and process.

When multiple equilibrium points can be Pareto ranked, it is possible to use concepts of efficiency to select a subset of self-enforcing equilibrium points: examples include R. Duncan Luce and Howard Raiffa's (1956, p. 106) concept of joint-admissibility, Tamer Basar and Geert Olsder's (1982, p. 72) concept of admissibility, and John Harsanyi and Reinhard Selten's (1988, p. 81) concept of payoff-dominance. An equilibrium point is said to be payoff-dominant if it is not strictly Pareto dominated by any other equilibrium point. When unique, considerations of efficiency may induce players to focus on and, hence, select the payoff-dominant equilibrium point, see Thomas Schelling (1980, p. 291).

In period game A, the equilibrium points are strictly Pareto ranked. Each player prefers a larger minimum. The only equilibrium point not Pareto dominated by any other equilibrium point is the n-tuple $(\bar{e},\dots,\bar{e})$: the payoff-dominant equilibrium point. Consequently, payoff-dominance selects the n-tuple $(\bar{e},\dots,\bar{e})$ in game A.

Selecting the unique payoff-dominant equilibrium point not only allows players to coordinate on an equilibrium point but also ensures that they will not coordinate on an inefficient one. Payoff dominance solves both the individual and the collective coordination problems of disequilibrium and coordination failure and, as Harsanyi and Selten suggest, *should* take precedence over alternative selection principles.

The tacit coordination game A provides a severe test of payoff dominance, because the minimum rule exacerbates the influence of uncertainty about the strategies of the other $n-1$ players. Define the cumulative distribution function for a player's action as $F(e_j)$. In the payoff-dominant equilibrium, $F(\bar{e})$ equals 1 and $F(e_j)$ equals 0 for e_j less than $\bar{e}$. A well-known theorem is that if $e_1,\dots,e_n$ are independent and identically distributed with common cumulative distribution function $F(e_j)$, then the cumulative distribution function for the minimum, $F_{\min}(e)$, equals $1-[1-F(e_j)]^n$; see A. M. Mood, F. A. Graybill, and D. C. Boes (1974). In the payoff-dominant equilibrium, $F_{\min}(\bar{e})$ equals 1 and $F_{\min}(e)$ equals 0 for e less than $\bar{e}$. But suppose that a player is uncertain that the $n-1$ players will select the payoff-dominant action, $\bar{e}$. Specifically, let $F(1)$ be small but greater than 0, then as n goes to infinity $F_{\min}(1)$ goes to 1. Consequently, when the number of players is large it only takes a remote possibility that an individual player will not select the payoff-dominant action $\bar{e}$ to motivate defection from the payoff-dominant equilibrium.

Several deductive selection principles based on the "riskiness" of an equilibrium point have been identified and formalized. A maximin action, which is an action (pure strategy) with the largest payoff in the worst possible outcome, is secure, see John Von Neumann and Oskar Morgenstern (1944, 1972). Given existence, security selects the equilibrium point supported by player's maximin actions. Security may select very inefficient equilibrium points in nonzero sum games.

In period game A, a player can ensure a payoff of $a-b$ by choosing e_i equal to 1, which is the largest payoff in the worst possi-

---

[2] The "disequilibrium Keynesians" emphasize the first coordination problem. The "equilibrium Keynesians" emphasize the second coordination problem, see Cooper and John (1988) for examples and references.

[3] We use the term induction in the logical, rather than mathematical, sense of reasoning from observed facts—history—to a conclusion.

TABLE 1—EXPERIMENTAL DESIGN

Experiment No.	Date	Size	*A* Payoff *A* Fullsize	*B* Payoff *B* Fullsize	*A′* Payoff *A* Fullsize	*C* Payoff *A* Size Two[a]
1	June	16	1^P,2,...,10	–	–	–
2	June	16	1^P,2,...,10^P	11,...,15	16^P,...,20	–
3	June	14	1^P,2,...,10^P	11,...,15	16^P,...,20	–
4	Sept	15	1^P,2^P,...,10^P	11^P,...,15	16,...,20	21,...,27
5	Sept	16	1^P,2^P,...,10^P	11^P,...,15	16,...,20	21,...,27
6	Sept	16	1^P,2^P,...,10^P	11^P,...,15	16,...,20	21,...,25
7	Sept	14	1^P,2^P,...,10^P	11^P,...,15	16,...,22	23,...,25

^{P} ~ Denotes a period in which subjects made predictions.
[a] ~ In experiment 4 and 5 pairings were fixed, while in experiments 6 and 7 pairings were random.

ble outcome. Consequently, in this tacit coordination game, security selects the n-tuple $(1,\ldots,1)$. Since payoff-dominance and security select different equilibrium points in this tacit coordination game (equilibrium points with the highest and lowest payoffs, respectively), an important and tractable empirical question is which, if any, deductive selection principle organizes the experimental data.

It is often possible to apply more than one deductive selection principle to a game. Hence, subjects may choose disequilibrium outcomes unless they behave as if there is a hierarchy of selection principles. When deductive selection principles fail to coordinate beliefs and actions, inductive selection principles based on repeated interaction may allow players to *learn to coordinate*.

Consider a finitely repeated game $A(T)$, which involves the n players playing period game A for T periods. The payoff-dominant equilibrium of $A(T)$ is just the repeated implementation of the payoff-dominant equilibrium of period game A, because the first-best outcome for period game A is $(\bar{e},\ldots,\bar{e})$. Similarly, the secure equilibrium of $A(T)$ is just the repeated implementation of the secure equilibrium of period game A.[4] Notice that the principles of efficiency and security can be defined independently of equilibrium. The experiments in this paper, which are designed to study the conflict between efficiency and security, are *not* designed to study how repeated play of period game A influences the set of equilibrium points for $A(T)$.

Having t periods of experience in $A(T)$ provides a player with observed facts, in addition to the description of the game, that can be used to reason about the equilibrium selection problem in the continuation game $A(T-t)$. This experience may influence the outcome of the continuation game $A(T-t)$ by focusing expectations on a specific equilibrium point. For example, one adaptive hypothesis is that players will give a best response to the minimum observed in the previous period. This adaptive behavior would immediately converge to an equilibrium in $A(T-1)$. The selected equilibrium involves all players choosing the period 1 minimum for the $T-1$ periods of the continuation game $A(T-1)$.

III. Experimental Design

Table 1 outlines the design of the seven experiments reported in this paper. The instructions were read aloud to ensure that the description of the game was common information, if not, common knowledge.[5] No pre-

---

[4] Crawford (1989) emphasizes that the secure equilibrium is also the only equilibrium that is evolutionary stable. In repeated play, players using adaptive behavior may be led to implement the secure equilibrium. Hence, while the experiments reported in this paper discriminate sharply between strategic stability and evolutionary stability, they do not discriminate between learning to use security and certain kinds of adaptive behavior.

[5] The original working paper, Van Huyck, Battalio, and Beil (1987), which includes the actual instructions,

PAYOFF TABLE *A*

		Smallest Value of X Chosen						
		7	6	5	4	3	2	1
Your	7	1.30	1.10	0.90	0.70	0.50	0.30	0.10
Choice	6	–	1.20	1.00	0.80	0.60	0.40	0.20
of	5	–	–	1.10	0.90	0.70	0.50	0.30
X	4	–	–	–	1.00	0.80	0.60	0.40
	3	–	–	–	–	0.90	0.70	0.50
	2	–	–	–	–	–	0.80	0.60
	1	–	–	–	–	–	–	0.70

PAYOFF TABLE *B*

		Smallest Value of X Chosen						
		7	6	5	4	3	2	1
Your	7	1.30	1.20	1.10	1.00	0.90	0.80	0.70
Choice	6	–	1.20	1.10	1.00	0.90	0.80	0.70
of	5	–	–	1.10	1.00	0.90	0.80	0.70
X	4	–	–	–	1.00	0.90	0.80	0.70
X	3	–	–	–	–	0.90	0.80	0.70
	2	–	–	–	–	–	0.80	0.70
	1	–	–	–	–	–	–	0.70

play negotiation was allowed. After each repetition of the period game, the minimum action was publicly announced and the subjects calculated their earnings for that period. The only common historical data available to the subjects was the minimum.

During the course of an experiment some design parameters were altered resulting in a sequence of treatments labeled *A*, *B*, *A'*, and *C*. Instructions for continuation treatments were given to the subjects after earlier treatments had been completed. The feasible actions in all treatments of all experiments were the integers 1 through 7: hence, $\bar{e}$ equaled 7.

In treatment *A* and *A'*, the following values were assigned to the parameters in equation (1): parameter a was set equal to \$0.20, parameter b was set equal to \$0.10, and a constant of \$0.60 was added to ensure that all payoffs were strictly positive. Consequently, the payoff-dominant equilibrium, $(7,\dots,7)$, paid \$1.30 while the secure equilibrium, $(1,\dots,1)$, paid \$0.70 per subject per period.[6] Subjects were given this information in the form of a payoff table, see payoff Table *A*. In treatment *A*, the period game *A* was repeated ten times. The number of players, n, varied between 14 and 16 subjects. (In treatment *C*, the number of players, n, was reduced to two.) Treatment *A'* designates the resumption of these conditions after treatment *B*.

In treatment *B*, parameter b in equation (1) was set equal to zero, see payoff Table *B*. This gives the subjects a dominating strategy (play 7 regardless of the minimum), which eliminates the coordination problem. The number of players, n, remained the same as in treatment *A*.

---

payoff tables, questionnaire, extra instructions and record sheet used in the experiments and a more extensive analysis of the experimental results, is available from the authors upon request.

[6] For the remainder of this paper, an equilibrium denotes a mutual best-response *outcome* in the period game.

Occasionally, subjects were asked to predict the actions of all the subjects in the treatment.[7] For each prediction in the September experiments, a subject was paid \$0.70 less 0.02 times the sum of the absolute value of the difference between the actual and predicted actions. (The rule used in the June experiments was less sensitive to prediction errors). At the end of the experiment, the subjects were told the actual distribution of actions and were paid.

The subjects were undergraduate students attending Texas A&M University and were recruited form sophomore and junior economics courses. A total of 107 students participated in the seven experiments. After reading the instructions, but before the experiment began, the students filled out a questionnaire to determine that they understood how to read the payoff table for treatment A, that is, map actions into money payoffs. The instructions would have been re-read if needed, but all 107 students responded correctly.

IV. Experimental Results

Table 2 reports the experimental results for treatment A. The data in period one are particularly interesting because the subjects can only use deductive selection principles to inform their behavior.

In period one, the payoff-dominant action, 7, was chosen by 31 percent of the subjects (33 of 107) and the secure action, 1, was chosen by 2 percent of the subjects (2 of 107). Neither deductive selection principle succeeds in organizing much of the data, although payoff-dominance is more successful than security. The popularity of actions 4 and 5—chosen by 18 and 34 subjects, respectively—is consistent with many subjects having nearly diffuse prior beliefs about the outcome of the period game.

The initial play of all seven experiments exhibit both individual and collective coordination failure. The minimum action for period one was never greater than 4. Hence, the largest payoff in period one was \$1.00 and some payoffs were \$0.10. (The payoff-dominant equilibrium would have paid everyone \$1.30). All of these outcomes are inefficient. The subjects were unable to use any deductive selection principle to coordinate on an equilibrium point.

Only 10 percent of the subjects predicted an equilibrium outcome in period one. Instead, most subjects (95 of 106) predicted a disequilibrium outcome.[8] Moreover, the subjects' predictions were dispersed: one third of the subjects predicted at least one 1 and one 7—a range of 6—and the average range of the predictions was 4.0. The subjects' dispersed predictions suggest that they expected other subjects to respond to the payoff table differently than they did. These data are inconsistent with any theory of equilibrium selection that assumes that, because a player will derive his prior probability distribution over other players' pure strategies strictly from the parameters of the game, all players will have the same prior probability distribution. Instead, some subjects made optimistic predictions and some subjects made pessimistic predictions.

While the subjects were unable to coordinate beliefs and actions, in almost all cases their individual predictions and actions were consistent. Of the 107 subjects, 106 subjects predicted that at least one other subject would choose an action equal to or less than their choice. (Only one subject predicted he would determine the minimum.) Most subjects mapped predictions into actions in a reasonable way. Those subjects who made pessimistic predictions about what the other subjects would do chose small values for their action and subjects who made optimistic predictions about what the other subjects would do chose large values for their action.[9]

---

[7] In two earlier pilot experiments predictions were not made in any period. The substantive results were the same as those reported here.

[8] One subject is excluded due to predicting only 15 choices.

[9] An anomaly is that, of the 95 subjects who predicted a minimum less than 7, 87 subjects chose an action greater than the minimum they predicted. Van Huyck, Battalio, and Beil (1987) provide an expected value model to explain this anomaly.

TABLE 2—EXPERIMENTAL RESULTS FOR TREATMENT *A*

	Period									
	1	2	3	4	5	6	7	8	9	10
Experiment 1										
No. of 7's	8	1	1	0	0	0	0	0	0	1
No. of 6's	3	2	1	0	0	0	0	0	0	0
No. of 5's	2	3	2	1	0	0	1	0	0	0
No. of 4's	1	6	5	4	1	1	1	0	0	0
No. of 3's	1	2	5	5	4	1	1	1	0	1
No. of 2's	1	2	2	4	8	7	8	6	4	1
No. of 1's	0	0	0	2	3	7	5	9	12	13
Minimum	2	2	2	1	1	1	1	1	1	1
Experiment 2										
No. of 7's	4	0	1	0	0	0	0	0	0	1
No. of 6's	1	0	1	0	0	1	0	0	0	0
No. of 5's	3	3	2	1	0	0	1	1	0	1
No. of 4's	4	6	2	3	3	0	0	0	0	0
No. of 3's	1	4	2	5	0	1	1	0	1	0
No. of 2's	3	2	6	5	5	9	3	4	3	1
No. of 1's	0	1	2	2	8	5	11	11	12	13
Minimum	2	1	1	1	1	1	1	1	1	1
Experiment 3										
No. of 7's	4	4	1	0	1	1	1	0	0	2
No. of 6's	2	0	2	0	0	0	0	0	0	0
No. of 5's	5	6	1	1	1	0	0	0	0	0
No. of 4's	3	3	2	1	2	1	0	0	0	1
No. of 3's	0	0	7	6	0	2	3	0	0	0
No. of 2's	0	1	1	4	5	3	6	3	2	2
No. of 1's	0	0	0	2	5	7	4	11	12	9
Minimum	4	2	2	1	1	1	1	1	1	1
Experiment 4										
No. of 7's	6	0	1	1	0	0	1	0	0	0
No. of 6's	0	6	2	0	0	1	0	0	0	0
No. of 5's	8	5	5	5	0	1	0	0	0	0
No. of 4's	1	1	4	6	7	1	2	1	1	0
No. of 3's	0	2	3	2	4	3	2	2	1	0
No. of 2's	0	1	0	0	2	3	7	4	2	2
No. of 1's	0	0	0	1	2	6	3	8	11	13
Minimum	4	2	3	1	1	1	1	1	1	1

An interesting question is whether the subjects' predictions correspond to the actual distribution of actions more closely than predictions based on payoff-dominance or security. Using the number of actions correctly predicted as a statistic, the data reveal that 95 percent of the subjects predicted the actions of the other $n-1$ subjects more accurately than did payoff-dominance. This statistic is used to measure prediction accuracy because the subjects payoff were a linear transformation of the prediction accuracy score. The difference of the mean prediction accuracy score was always positive and in most cases significantly different from zero at the 1 percent level: a result that is robust to non-parametric statistical procedures. Obviously, security does even worse. Subjects predicted the observed heterogenous response to the description of the game and the resulting coordination failure in period one.

Repeated play of the period game allows the subjects to use inductive selection principles or to learn to use deductive selection principles. Hence, repeated play makes it more likely that subjects will be able to obtain mutual best-response outcomes in the continuation game. Repeating the period game does cause actions to converge to a

Table 2—Experimental Results for Treatment *A*, Continued

	Period									
	1	2	3	4	5	6	7	8	9	10
Experiment 5										
No. of 7's	2	2	3	1	1	1	1	0	0	0
No. of 6's	1	3	1	0	0	0	0	0	0	0
No. of 5's	9	3	0	4	1	0	2	0	0	0
No. of 4's	3	4	6	2	1	2	0	2	1	1
No. of 3's	1	2	2	4	6	0	0	0	0	1
No. of 2's	0	2	2	3	4	6	5	2	5	3
No. of 1's	0	0	2	2	3	7	8	12	10	11
Minimum	3	2	1	1	1	1	1	1	1	1
Experiment 6										
No. of 7's	5	3	1	1	1	1	2	2	2	3
No. of 6's	2	0	0	0	1	0	0	0	0	0
No. of 5's	5	1	0	0	0	1	0	0	0	0
No. of 4's	2	3	4	0	0	0	0	0	0	0
No. of 3's	1	5	4	2	2	2	1	0	2	0
No. of 2's	0	2	4	5	3	3	6	4	5	5
No. of 1's	1	2	3	8	9	9	7	10	7	8
Minimum	1	1	1	1	1	1	1	1	1	1
Experiment 7										
No. of 7's	4	3	1	1	1	1	1	1	1	1
No. of 6's	1	0	0	0	0	0	0	0	0	0
No. of 5's	2	3	0	0	0	0	0	0	0	0
No. of 4's	4	0	1	2	1	0	0	0	0	0
No. of 3's	1	3	2	1	1	0	0	0	0	0
No. of 2's	1	3	2	2	4	4	4	4	5	3
No. of 1's	1	2	8	8	7	9	9	9	8	10
Minimum	1	1	1	1	1	1	1	1	1	1

stable outcome, see Table 2. But rather than converging to the payoff-dominant equilibrium or to the initial outcome of the treatment, the most inefficient outcome obtains in all seven experiments.

The change in a subject's action between period one and period two provides insight into the subjects' dynamic behavior. Of the eleven subjects who determined the minimum in period one the average change in action between period one and two was 0.73: seven subjects increased their action, three did not change, and one decreased his action. In every experiment someone who had not determined the minimum in period one determines the minimum in period two. Moreover, in experiments one through five the intersection of the set of subjects who determine the minimum in period one with the set of subjects who determine the minimum in period two is empty. Since a subject's payoff is increasing in e_i when he (she) uniquely determines the minimum, this adaptive behavior can be rationalized.

However, a subject's payoff is decreasing in e_i when he (she) played above the minimum and those subjects that played above the minimum reduced their choice of action. The observed mean reduction is increasing in the difference between a subject's action and the reported minimum, and the mean reduction is smaller than this difference. The observed correlation between the current choice of these optimistic subjects and the minimum reported in period 1 suggests that behavior in the continuation game $A(9)$ is not independent of the history leading up to continuation game $A(9)$. However, only 14 of 107 subjects give a best-response to the period one minimum in period two.

Some subjects play below the minimum of the preceding period. This observed "overshooting" cannot be reconciled with adaptive theories that predict the current action will be a convex combination of last periods action and last periods outcome. Apparently, some subjects learn how "risky" it is to choose an action other than the secure action, 1, under the minimum rule and learn to use security to inform their behavior in the continuation game.

Although it failed to predict the initial outcome, security predicts the stable outcome of period game A. By period ten 72 percent of the subjects (77 out of 107) adopt their secure action, 1, and the minimum for all seven experiments was a 1. The observed coordination failure appears to result from a few subjects concluding it is too "risky" to choose an action other than the secure action and from most subjects focusing on the minimum reported in earlier period games. The minimum rule interacting with this dynamic behavior causes this treatment to converge to the most inefficient outcome.

In treatment B, parameter b of equation (1) was set equal to zero. Because a player's action is no longer penalized, the payoff-dominant action, 7, is a best response to all feasible minimums. Action 7 is a dominating strategy. Hence, treatment B tests equilibrium *refinements* based on the elimination of individually unreasonable actions. For example, a simple dominance argument eliminates all of the equilibrium points except one: $(7,\ldots,7)$. Any strategic uncertainty would cause an individually rational player to choose the payoff-dominant action, 7.

Table 3 reports the experimental results for treatment B and treatment A'. In period eleven, the payoff-dominant action, 7, was chosen by 84 percent of the subjects (76 of 91). However, the minimum in period eleven was never more than 4 and in experiments four, five, six, and seven it was a 1.[10]

Of course, a subject that adopts action 7 need not worry about what actions other subjects take and, apparently, most subjects did not. This property of dominating strategies resulted in the B treatment exhibiting different dynamics than the A treatment. Like the A treatment, those players who determine the minimum increase their action, but, unlike the A treatment, those players who were above the minimum do not decrease their action.[11] This dynamic behavior converges to the efficient outcome—the payoff-dominant equilibrium—in four of the six experiments. By period fifteen, 96 percent of the subjects chose the payoff-dominant action, 7.

Even in the experiments that obtained the efficient outcome, the B treatment was not sufficient to induce the groups to implement the payoff-dominant equilibrium in treatment A': parameter b equals \$0.10 once again. Returning to the original payoff table in period sixteen, 25 percent of the subjects chose the payoff-dominant action, 7.[12] However, 37 percent chose the secure action, 1. Period sixteen predictions were peaked with subjects choosing a 7 predicting most subjects would choose 7 and subjects choosing a 1 predicting most subjects would choose 1. This bi-modal distribution of actions and predictions suggests that play prior to period sixteen influenced subjects' behavior. However, the subjects exhibit a heterogenous response to this history.

Security predicts the stable outcome of treatment A'. In treatment A', the minimum in all periods of all six experiments was 1. By period twenty, 84 percent of the subjects chose the secure action, 1, and 94 percent chose an action less than or equal to 2. (Experiments two and four even satisfy the

---

[10] At least one subject did not understand how the payoff table had changed. Subject 3 in experiment 5, who plays a 1 in every period of the B treatment, predicts that all 16 players will choose 1 but only he does so. When the actual distribution was revealed, subject 3 appeared genuinely amazed and confessed that he had not understood how the payoffs had changed.

[11] The two exceptions were due to subject 3 in experiment five, see fn. 10, and subject 12 in experiment six. Subject 12 predicts that he will uniquely determine the minimum in period 11, verifies this in period 12 by choosing a 3, and then chooses a one for the remainder of the B treatment. Perhaps, subject 12 became vindictive. He had chosen a 7 in period 1.

[12] The large fluctuations in behavior resulting from changes in the parameter b—between treatment A and B and between treatment B and A'—suggest that subjects are influenced by the description of the game. In our view, these data are inconsistent with backward-looking theories of adaptive behavior.

TABLE 3—EXPERIMENTAL RESULTS FOR TREATMENT B AND TREATMENT A'

	Treatment B					Treatment A'				
	11	12	13	14	15	16	17	18	19	20
Experiment 2										
No. of 7's	13	15	16	16	16	8	2	0	0	0
No. of 6's	1	0	0	0	0	0	0	0	0	0
No. of 5's	0	1	0	0	0	1	0	0	0	0
No. of 4's	1	0	0	0	0	1	2	0	0	0
No. of 3's	1	0	0	0	0	1	1	1	1	0
No. of 2's	0	0	0	0	0	3	3	4	2	0
No. of 1's	0	0	0	0	0	2	8	11	13	16
Minimum	3	5	7*	7*	7*	1	1	1	1	1*
Experiment 3										
No. of 7's	13	13	12	13	14	6	2	2	1	1
No. of 6's	0	0	1	1	0	1	0	0	0	0
No. of 5's	0	0	1	0	0	0	2	1	0	0
No. of 4's	1	0	0	0	0	1	0	0	0	1
No. of 3's	0	1	0	0	0	0	0	0	0	0
No. of 2's	0	0	0	0	0	2	4	2	3	0
No. of 1's	0	0	0	0	0	4	6	9	10	12
Minimum	4	3	5	6	7*	1	1	1	1	1
Experiment 4										
No. of 7's	12	13	14	14	15	3	1	0	0	0
No. of 6's	0	0	0	0	0	0	0	0	0	0
No. of 5's	1	0	0	1	0	0	0	0	0	0
No. of 4's	0	1	1	0	0	2	0	0	0	0
No. of 3's	0	1	0	0	0	2	0	0	0	0
No. of 2's	0	0	0	0	0	2	1	2	0	0
No. of 1's	2	0	0	0	0	6	13	13	15	15
Minimum	1	3	4	5	7*	1	1	1	1*	1*
Experiment 5										
No. of 7's	13	13	15	15	15	1	0	0	0	0
No. of 6's	0	0	0	0	0	0	0	0	0	0
No. of 5's	1	1	0	0	0	0	0	0	0	0
No. of 4's	1	1	0	0	0	0	0	0	0	0
No. of 3's	0	0	0	0	0	1	1	0	0	0
No. of 2's	0	0	0	0	0	3	4	2	2	3
No. of 1's	1	1	1	1	1	11	11	14	14	13
Minimum	1	1	1	1	1	1	1	1	1	1
Experiment 6										
No. of 7's	13	13	12	12	13	2	2	2	2	2
No. of 6's	0	1	1	1	0	0	0	0	0	0
No. of 5's	0	1	1	0	1	0	0	0	0	0
No. of 4's	1	0	1	1	0	1	0	0	0	0
No. of 3's	0	1	0	1	0	1	0	0	0	0
No. of 2's	1	0	0	0	1	5	6	7	6	5
No. of 1's	1	0	1	1	1	7	8	7	8	9
Minimum	1	3	1	1	1	1	1	1	1	1
Experiment 7										
No. of 7's	12	14	13	13	14	3	4	2	2	2
No. of 6's	0	0	1	0	0	0	0	0	0	0
No. of 5's	0	0	0	0	0	1	0	0	0	0
No. of 4's	1	0	0	0	0	2	0	0	0	0
No. of 3's	0	0	0	1	0	2	0	0	0	0
No. of 2's	0	0	0	0	0	2	4	2	2	1
No. of 1's	1	0	0	0	0	4	6	10	10	11
Minimum	1	7*	6	3	7*	1	1	1	1	1

* ~ Denotes a mutual best-response outcome.

mutual best-response property of an equilibrium by period twenty.) Obtaining the efficient outcome in treatment *B* failed to reverse the observed coordination failure. Like the *A* treatment, the most inefficient outcome obtained.

V. Experimental Results for Treatment *C*: Group Size Two

Treatment *C* was added to the September experiments to determine if subjects were influenced by group size when choosing their actions. In theory, any uncertainty about the actions of an individual player in the game is exacerbated by the minimum rule as the number of players increases, see Section II. The *C* treatment reduces group size to two.

Table 4 reports the experimental results for the *C* treatment of experiments four and five, which permanently paired subjects with an unknown partner. In period twenty-one, 42 percent of the subjects play their payoff-dominant action, 7, and 74 percent of the subjects increase their action. This result occurred even though the minimum for the preceding five periods had been a 1 and all 31 subjects had played either a 1 (28 subjects) or a 2 (3 subjects) in period twenty. Clearly, either the subjects thought that their partner in treatment *C* would change his (her) action in response to reduced group size or the subjects expected alternative dynamics in repeated play.

The subjects in experiments four and five used an adaptive behavior in the *C* treatment similar to the adaptive behavior exhibited in the *A* treatment. Subjects that played the minimum increased effort by an average of +2.0 and subjects that played above the minimum reduced effort by an average of −1.9. However, unlike the *A* treatment, there was no "overshooting" to the secure action, 1. Occasionally, both subjects simultaneously chose the payoff-dominant action, 7.[13] This dynamic behavior converged to the efficient outcome—the payoff-dominant equilibrium—in 12 of 14 trails. Hence, even with an extremely negative history payoff-dominance predicts the stable outcome of the tacit coordination game with fixed pairs.

Experiments six and seven randomly paired subjects with an unknown partner.[14] Hence, experiments six and seven test whether the results obtained in Experiments four and five were due to subjects repeating the period game with the same opponent. In the first period of these experiments, 37 percent of the subjects chose the payoff-dominant action, 7, and 73 percent of the subjects increased their choice of action, see Table 5. Moreover, the subjects' dynamic behavior was similar to that found in the fixed pair *C* treatment. While the results for the random pair *C* treatment are influenced by group size, no stable outcome obtains.

The *C* treatment confirms that there are two consequences of the minimum rule. First, group size interacting with the minimum rule alters the subjects' initial choice of action. Second, group size interacting with the minimum rule alters the convergence of the subjects' dynamic behavior in disequilibrium.

---

[13] Recall that subjects only observe the minimum and their own action. Hence, it is not possible to unilaterally "signal" a willingness to implement the payoff-dominant equilibrium, that is, subjects could not use Osborne's (1987) refinement of a "convincing deviation" to inform their behavior, see also van Damme (1987).

[14] Experiments by Cooper, DeJong, Forsythe, and Ross (1987) report that after eleven repetitions randomly paired groups of size two almost always obtain the payoff-dominant equilibrium. However, Cooper et al. also report coordination failure when subjects can choose from a strategy space that includes certain kinds of dominated cooperative strategies. Their game illustrates an interesting distinction between Luce and Raiffa's "solution in the strict sense," which depends on joint-admissibility and Harsanyi and Selten's solution, which depends on payoff-dominance. Because the first-best outcome requires using a strictly dominated strategy—as in the prisoner's dilemma game—and because joint-admissibility admits efficiency comparisons with disequilibrium outcomes, the Cooper et al. game with dominated cooperative strategies has no "solution in the strict sense" of Luce and Raiffa. Because the first-best outcome is an equilibrium in period game *A*, joint-admissibility—appropriately defined for n person games—and payoff-dominance select the same equilibrium point in period game *A*.

TABLE 4—EXPERIMENTAL RESULTS FOR TREATMENT *C*: FIXED PAIRINGS

	Period						
	21	22	23	24	25	26	27
Experiment 5							
Pair 1							
Subject 1	7	7	7	7	7	7	7
Subject 16	7	7	7	7	7	7	7
Minimum	7*	7*	7*	7*	7*	7*	7*
Pair 2							
Subject 2	7	2	7	7	7	7	7
Subject 15	1	7	3	6	7	7	7
Minimum	1	2	7	7	7	7	7
Pair 3							
Subject 3	1	1	1	1	1	1	1
Subject 14	1	1	7	1	1	1	7
Minimum	1*	1*	1	1*	1*	1*	1
Pair 4							
Subject 4	1	7	7	7	7	7	7
Subject 13	7	2	5	7	7	7	7
Minimum	1	2	5	7*	7*	7*	7*
Pair 5							
Subject 5	1	7	4	7	7	7	7
Subject 12	1	4	7	7	7	7	7
Minimum	1	4	4	7*	7*	7*	7*
Pair 6							
Subject 6	5	7	7	7	7	7	7
Subject 11	7	7	7	7	7	7	7
Minimum	5	7*	7*	7*	7*	7*	7*
Pair 7							
Subject 7	1	7	6	7	7	7	7
Subject 10	5	3	6	7	7	7	7
Minimum	1	3	6*	7*	7*	7*	7*
Pair 8							
Subject 8	7	6	6	7	7	7	7
Subject 9	3	5	7	7	7	7	7
Minimum	3	5	6	7*	7*	7*	7*
Experiment 6							
Pair 1							
Subject 2	7	7	4	5	6	6	7
Subject 15	2	3	6	6	7	7	7
Minimum	2	3	4	5	6	6	7*
Pair 2							
Subject 3	5	7	7	7	7	7	7
Subject 14	7	7	7	7	7	7	7
Minimum	5	7*	7*	7*	7*	7*	7*
Pair 3							
Subject 4	1	1	1	1	4	4	1
Subject 13	7	1	1	3	1	1	2
Minimum	1	1*	1*	1	1	1	1
Pair 4							
Subject 5	5	7	7	7	7	7	7
Subject 12	7	7	7	7	7	7	7
Minimum	5	7*	7*	7*	7*	7*	7*

TABLE 4—FIXED PAIRINGS, Continued

	Period						
	21	22	23	24	25	26	27
Pair 5							
Subject 6	4	5	7	7	7	7	7
Subject 11	4	5	7	7	7	7	7
Minimum	4*	5*	7*	7*	7*	7*	7*
Pair 6							
Subject 7	5	7	7	7	7	7	7
Subject 10	5	7	7	7	7	7	7
Minimum	5*	7*	7*	7*	7*	7*	7*

* ~ Denotes a mutual best-response outcome.

TABLE 5—DISTRIBUTION OF ACTIONS FOR TREATMENT *C*: RANDOM PAIRINGS

	Period				
	21	22	23	24	25
Experiment 6					
No. of 7's	5	5	4	10	8
No. of 6's	0	1	3	0	0
No. of 5's	2	5	3	3	4
No. of 4's	3	1	1	1	1
No. of 3's	1	1	1	0	0
No. of 2's	1	1	2	2	2
No. of 1's	4	2	2	0	1
Experiment 7					
No. of 7's	–	–	6	5	5
No. of 6's	–	–	1	0	1
No. of 5's	–	–	0	3	0
No. of 4's	–	–	2	1	4
No. of 3's	–	–	2	0	0
No. of 2's	–	–	0	0	1
No. of 1's	–	–	3	5	3

VI. Treatment A with Monitoring

As a referee points out, a reasonable conjecture is that revealing the distribution of actions each period—in addition to the minimum—might influence the reported dynamics. For example, subjects could signal a willingness to coordinate on the payoff-dominant equilibrium and optimistic subjects might delay reducing their action if they knew the minimum was determined by just one subject. Two experiments, each using payoff Table *A* and 16 naive subjects, were conducted in which the entire distribution of actions was recorded on a blackboard at the end of each period and was left there for the entire experiment.

The initial distribution of actions and the dynamics of the two monitoring experiments were similar to those reported above, see Table 6. If anything, the convergence of actions to the secure action, 1, was more rapid under the monitoring treatment. In fact, a mutual best-response outcome was obtained in one experiment: without monitoring mutual best-response outcomes are not observed for period game *A* until treatment *A'*. Apparently, monitoring helps solve the individual coordination problem—more subjects give a best-response sooner—but

TABLE 6—DISTRIBUTION OF ACTIONS FOR TREATMENT *A* WITH MONITORING

	Period									
	1^p	2^p	3	4	5	6	7	8	9	10^p
Experiment 8										
No. of 7's	4	0	0	0	2	0	0	1	1	1
No. of 6's	1	1	0	0	0	0	0	0	0	0
No. of 5's	4	0	1	0	0	1	0	0	0	1
No. of 4's	5	4	2	1	0	0	1	0	0	0
No. of 3's	1	4	1	0	0	2	0	0	0	0
No. of 2's	0	2	1	2	3	2	1	1	2	1
No. of 1's	1	5	11	13	11	11	14	14	13	13
Minimum	1	1	1	1	1	1	1	1	1	1
Experiment 9										
No. of 7's	6	2	0	1	0	0	0	0	0	0
No. of 6's	1	2	0	0	0	0	0	0	0	0
No. of 5's	2	2	1	0	0	0	0	1	0	0
No. of 4's	4	2	3	1	0	0	0	0	0	1
No. of 3's	1	5	3	1	0	0	0	0	1	0
No. of 2's	0	1	0	5	4	1	0	0	1	2
No. of 1's	2	2	9	8	12	15	16	15	14	13
Minimum	1	1	1	1	1	1	1*	1	1	1

^{p} ~ Denotes a period in which subjects made predictions.
* ~ Denotes a mutual best-response outcome.

not the collective coordination problem—the minimum was a 1 in all ten periods of both experiments.

VII. Concluding Comments

These experiments provide an interesting example of coordination failure. The minimum was never above four in period one and all seven experiments converged to a minimum of one within four periods. Since the payoff-dominant equilibrium would have paid all subjects \$19.50 in the *A* and *A′* treatments—excluding predictions—and the average earnings were only \$8.80, the observed behavior cost the average subject \$10.70 in lost earnings.

This inefficient outcome is not due to conflicting objectives as in "prisoner's dilemma" games or to asymmetric information as in "moral hazard" games. Rather, coordination failure results from *strategic uncertainty*: some subjects conclude that it is too "risky" to choose the payoff-dominant action and most subjects focus on outcomes in earlier period games. The minimum rule interacting with this dynamic behavior causes the *A* and *A′* treatments to converge to the most inefficient outcome.

Deductive methods imply that all feasible actions are consistent with some equilibrium point in this experimental coordination game. However, the experimental results suggest that the first-best outcome, which is the payoff-dominant equilibrium, is an extremely unlikely outcome either initially or in repeated play. Instead, the results suggest that the initial outcome will not be an equilibrium point and only the secure—but very inefficient—equilibrium describes behavior that actual subjects are likely to coordinate on in repeated play of period game *A* when the number of players is not small.

REFERENCES

Basar, Tamer and Olsder, Geert J., *Dynamic Noncooperative Game Theory*, New York: Academic Press, 1982.

Bryant, John, "A Simple Rational Expectations Keynes-Type Model," *Quarterly Journal of Economics*, August 1983, *98*, no. 3, 525–28.

Cooper, Russell, DeJong, Douglas V., Forsythe, Robert and Ross, Thomas W., "Selection Criteria in Coordination Games: Some Experimental Results," unpublished manuscript, October 1987.

______ **and John, Andrew**, "Coordinating Coordination Failures in Keynesian Models," *Quarterly Journal of Economics*, August 1988, *103*, no. 3, 441–64.

Crawford, Vincent P., "An 'Evolutionary' Interpretation of Van Huyck, Battalio, and Beil's Experimental Results on Coordination," unpublished manuscript, January 1989.

Harsanyi, John C., and Selten, Reinhard, *A General Theory of Equilibrium Selection in Games*, Cambridge: MIT Press, 1988.

Kohlberg, Elon and Mertens, Jean-Francois, "On the Strategic Stability of Equilibria," *Econometrica*, *54*, no. 5, September 1986.

Lucas, Robert E., Jr., "Adaptive Behavior and Economic Theory," in *Rational Choice*, R. Hogarth and M. Reder, eds. Chicago: University of Chicago Press, 1987.

Luce, R. Duncan and Raiffa, Howard, *Games and Decisions*, New York: Wiley & Sons, 1957.

Mood, A. M., Graybill, F. A. and Boes, D. C., *Introduction to the Theory of Statistics*, 3rd ed., New York: McGraw-Hill, 1974.

Osborne, Martin J., "Signaling, Forward Induction, and Stability in Finitely Repeated Games," unpublished manuscript, November 1987.

Schelling, Thomas C., *The Strategy of Conflict*, Cambridge: Harvard University Press, 1980.

van Damme, Eric, "Stable Equilibria and Forward Induction," Universität Bonn Discussion Paper No. 128, August 1987.

Van Huyck, John B., Battalio, Raymond C. and Beil, Richard O., "Keynesian Coordination Games, Strategic Uncertainty, and Coordination Failure," unpublished manuscript, October 1987.

Von Neumann, John and Morgenstern, Oskar, *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1972.

Collusion Through Insurance: Sharing the Cost of Oil Spill Cleanups

By EDDIE DEKEL AND SUZANNE SCOTCHMER*

When the Exxon Valdez snagged on an underwater mountain in March 1989, and released 11 million gallons of crude oil into Prince William Sound, part of the public outrage was due to the subsequent rise in gasoline prices. Indeed, it seemed possible that, rather than being punished, the oil industry as a whole would profit from the oil spill, and this raised questions about the industry's incentives for taking care.[1] On the other hand, it was not obvious that Exxon itself would profit from the spill, since Exxon would be responsible for much of the cleanup costs, and, in addition, its flow of oil, hence revenue, would be disrupted. While Exxon might be careless in order to earn high profits at inflated prices, it certainly would not do so to confer profits on its competitors. Or would it?

Spills disrupt supply and increase prices. Even if a spill imposes a loss on the spiller, the disruption could increase the other firms' profits enough to more than compensate.[2] A cynic might conjecture that, if the oil companies had a credible way to collude in the amount of care they took, they might reduce care and increase the frequency of spills so as to increase their joint profit. If spills were distributed randomly among firms, all firms would profit on average. Of course, an individual firm could do even better by never spilling, and benefiting from other firms' spills through the higher price. Collusion to reduce the level of care would be unstable because each firm would have an incentive to increase its level of care. The necessary ingredient for a "conspiracy theory" is that firms can commit to reduce their levels of care. We argue that sharing the cost of cleanup serves exactly this function.

Since the oil industry maintains common resources to clean up oil spills,[3] the cost of any member's spill is shared by all of the members. Since spills are random, this cost sharing is a form of risk sharing. (Cost sharing implies that a firm pays less in the event that it alone has a spill, and more when another firm has a spill, than without cost sharing.) Risk sharing should lead to moral hazard. Here, the moral hazard "problem" is that each firm will reduce its level of care relative to the amount of care it would take if it bore the entire cost of its own cleanup. When agents are risk averse, the benefits of risk sharing can offset the costs of moral hazard. Here, however, it is hard to see any direct benefits of risk sharing, since one might expect diversely held corporations to be risk neutral.[4]

---

*Department of Economics, and Graduate School of Public Policy, University of California, Berkeley, CA 94720. We are grateful to Andreu Mas-Colell and Tom Romer for useful comments, and to the NSF, grant nos. SES-8808133 and 8809107, and to U.C., Berkeley, Committee on Research, for financial support.

[1] See Ann Crittendon, April 14, 1989, *Los Angeles Times*, p. 14, and Charles Lave and John Quigley (1989). The latter authors estimated that the oil industry would earn $100 million in additional revenues on the West Coast alone if the 10-percent rise in gasoline prices persisted for one month. By now, a year after the spill, the cleanup costs and legal liabilities threaten to be sufficiently large that the industry as a whole will not profit.

[2] Since oil is an exhaustible resource, a supply reduction today increases the reserves available tomorrow. This intertemporal substitution can, but need not, increase the present discounted value of industry profit. When deferring sales to the future reduces the aggregate industry profit, our arguments below do not apply.

[3] The oil consortium Alyeska owns the Alaska pipeline and is responsible for cleaning up its members' spills. Subsequent to three additional oil spills in the spring of 1989, the American Petroleum Institute recommended that the oil industry jointly finance several other oil spill cleanup centers on both coasts.

[4] Furthermore, spills are typically small relative to the oil companies' net worth. Exxon's cleanup costs are

Since risk sharing does not seem, on *prima facie* grounds, to be very important, we are motivated to look for a strategic reason to share cleanup costs. If firms in an oligopoly could collude, they would typically cut supply. They cannot collude directly, since that would violate antitrust laws. However, oil spills disrupt supply, and can therefore be profitable, just as if the firms had agreed to withhold supply to raise the price. Collusion can be enforced by sharing cleanup costs. When cleanup costs are shared, each firm has less incentive to take care.

Committing to low levels of care (hence, supply disruptions) by sharing cleanup costs is an example of how firms can create incentives to behave collusively without actually violating antitrust laws and without having to monitor each other. Other ways that firms can create incentives to collude are discussed by Steven Salop (1986). For example, M. Maloney, R. McCormick, and R. Tollison (1979) suggest that unionization (which leads to supply disruptions during strikes) can be used to increase industry profits; and Timothy Bresnahan and Steven Salop (1986), and Robert Reynolds and Bruce Snapp (1986) show how joint ventures can be used by parent firms to enhance collusion.[5]

I. Model

For ease of modeling, we make the following assumptions. We will use the term "profit" to refer to a firm's net revenue, excluding costs of cleanup and of care. For simplicity, we assume there are two firms. If there is no spill,then both firms produce oil and each firm's profit is π_2. If one firm has a spill, then its supply is disrupted,[6] so that it loses all profits, and the other firm makes larger profits, denoted by π_1. If both firms have spills, neither produces and, hence, per firm profit is zero. Oil spills are all the same size and each costs C to clean up, with no residual environmental damage. The cost of taking care level e is $c(e)$, and the probability of not having a spill is $p(e)$.

If there were no supply disruption, each firm's profit would be π_2 whether or not there were spills. The industry's profit function would be $2\pi_2-2c(e)-2(1-p(e))C$, and one firm's profit function would be $\pi_2-c(e)-(1-p(e))C$. There would be no divergence between the effort level that maximizes industry profit and the effort levels exerted in equilibrium. In fact, the firms' preferred levels of effort would be efficient, as one might expect since each firm bears all the costs of care and of cleaning up its spills.

Assuming now that spills lead to supply disruptions, and assuming that both firms take the same level of care e, the industry's profit, as a function of e, is

$$(1)\quad 2\pi_1 p(e)(1-p(e))+2\pi_2 p(e)^2 -2c(e)-2(1-p(e))C.$$

The derivative of the industry's profit function is

$$(2)\quad [\pi_1+C+(\pi_2-\pi_1)2p(e)] \times p'(e)-c'(e).$$

If firm 1 takes the level of care e and firm 2 takes the level of care e^*, then firm 1's expected profit, as a function of e, is

$$(3)\quad \pi_1 p(e)(1-p(e^*))+\pi_2 p(e)p(e^*) -c(e)-(1-p(e))C.$$

---

now about one billion dollars. Exxon's profits (not assets) were in the neighborhood of $5 billion in 1987.

[5] Another argument that the oil consortium Alyeska is strategic, rather than naive, is given by Richard Levin, Sharon Oster, and Steven Salop (1989) who argue that Alyeska's profit-sharing agreement encourages the members to price at the monopoly price.

[6] The supply disruption due to an oil spill can take many forms. In Alaska, it was due to the loss of oil in the tanker itself, and to the fact that flow through the pipeline had to be curtailed, since resources were being devoted to the cleanup. The consumer costs of the supply disruption will partly be smoothed through inventories, and will be shared by many firms.

In a symmetric Nash equilibrium, in which firm 1 undertakes the same level of effort as firm 2, the first-order condition that describes that effort level is[7]

$$(4) \qquad [\pi_1 + C + (\pi_2 - \pi_1) p(e)] \\ p'(e) - c'(e) = 0.$$

We assumed above that $\pi_2 < \pi_1$. Therefore, at the equilibrium effort levels that solve (4), the derivative of industry profit (2) is negative. A decrease in each firm's effort level would increase industry profit.

There exists $\alpha < 1$ such that, if we substitute αC for C in expression (4), industry profit is greater in Nash equilibrium than with $\alpha = 1$. Collusion has been achieved through risk sharing. (The spiller pays αC and the nonspiller pays $(1-\alpha)C$.) But, of course, the greater industry profit has been achieved through a reduction in care and an increase in spills.

Two simplifications in this model are (i) that the costs of care $c(e)$ are separable from "profit" and (ii) that the profit of a spiller in any period is zero. Regarding (i), instead of modeling the cost of effort as a separable cost function $c(e)$, we could let the monopoly and duopoly profits depend on e, as when care increases the unit cost of producing oil. In that case the terms $-c'(e)$ in the first-order conditions (2) and (4) must be replaced by $\pi_1'(e)p(e)[1-p(e)] + \pi_2'(e)p(e)^2$, and the same argument applies: At the symmetric equilibrium effort levels, the derivative of the industry's profit function is negative. Regarding (ii), a complete model would specify a profit level for each firm, and for each possible contingency. Denote by $\pi(i, j), i, j \in \{\text{spill, no spill}\}$, the profit of a firm with realization i (spill or no spill) when the other firm has realization j (spill or no spill). In this more general model, for our results to hold it is sufficient that $\pi(\text{spill,spill}) \geq 2\pi(\text{spill, no spill})$. That is, the firm earns significantly less profits if it is the only spiller than if both firms spill. This condition is satisfied by our previous simplifying assumption that $\pi(\text{spill, spill}) = \pi(\text{spill, no spill}) = 0$.

II. Remarks on Social Welfare

The argument above showed that one consequence of risk sharing is that it facilitates collusion by encouraging spills, reducing output, and, hence, increasing profits in equilibrium. Whether social welfare is enhanced or undermined by this collusion depends on whether a social planner would prefer more or less care than firms take in equilibrium. This, in turn, depends on aspects of the market (elasticity of demand, the cost of care, and the technology by which care translates into a reduced probability of spilling) and on whether there are additional externalities, such as residual damage that cannot be cleaned up.

The welfare comparison is ambiguous even if we focus on market issues and ignore additional externalities such as arise from incomplete cleanups. We will compare the equilibrium levels of effort with the effort level that a social planner would choose. The effort levels chosen by the planner determine the probabilities of spills, and, hence the probabilities that there are zero, one or two firms active in a given period. We assume the planner cannot control firms' oil supplies once the level of care is determined.

The welfare comparison can go either way. To see the intuition for this, assume that firm 2 has a fixed level of effort, and consider firm 1's incentive to increase care versus the social planner's. Firm 1 cares only about the profits that care provides. These profits result from being in the market when it does not have a spill and they may be monopoly or duopoly profits, according to whether the other firm spills. In contrast, the social planner cares about the total social surplus (consumers' surplus plus profits) which arise from firm 1 taking care and being in the market. In the case where firm 2 has a spill and is out of the market, the social planner sees a larger total return to care by firm 1 because the planner cares about the total surplus rather than only about profit. In the case where firm 2 does not have a spill and is in the market, firm 1 may value care more than or less than the social planner. On the one hand, firm 1 receives a transfer of profit from firm 2 if firm 1 does not have a spill, (firm 2 has less output and less profit as a duopolist than it would have as a monopolist) but the social planner does not value this transfer. On the other hand, the social planner does value the additional social surplus that expanded output can generate. Which of these considerations dominates will determine whether the social planner or firms in Nash equilibrium prefer higher effort levels. For example, if the profit level π_2 is determined by Bertrand competition (so that firms make zero profits if neither has a spill) then the levels of care in the symmetric Nash equilibrium are too low even without cost-sharing agreements, and will be lower still with such agreements.

---

[7] We are focusing on symmetric Nash equilibrium, although there may be other Nash equilibria. Our arguments apply if there is an interior symmetric Nash equilibrium ($0 < p(e) < 1$), as there may be if $p(e)$ and $-c(e)$ are concave.

The observation that oil companies might collude to increase profit through risk sharing is rather cynical. Some people may find it hard to believe that oil companies orchestrate spills so as to increase profit. Indeed they may not...but, nevertheless, cost sharing can decrease incentives for care and thereby increase profit.

REFERENCES

Bresnahan, Timothy and Salop, Steven, "Quantifying the Competitive Effects of Production Joint Ventures," *International Journal of Industrial Organization*, June 1986, *4*, 155–75.

Lave, Charles and Quigley, John, "Why is this Man Smiling?," mimeo., Graduate School of Public Policy, University of California, Berkeley, April 1989.

Levin, Richard, Oster, Sharon and Salop, Steven, "The Alaska Pipeline and Facilitating Devices: How Eight Competitors Can Be Worse Than One," mimeo., Yale School of Management, August 1989.

Maloney, Michael, McCormick, R. and Tollison, R., "Achieving Cartel Profits through Unionization," *Southern Economic Journal*, October 1979, *46*, 628–34.

Reynolds, Robert J. and Snapp, Bruce R., "The Competitive Effect of Partial Equity Interests and Joint Ventures," *International Journal of Industrial Organization*, June 1986, *4*, 141–53.

Salop, Steven, "Practices that (Credibly) Facilitate Tacit Coordination," in G. F. Matthewson and J. Stiglitz, eds., *New Developments in the Analysis of Market Structure: Proceedings of a Conference Held by the International Economic Association in Ottawa, Canada*, Cambridge, MA: MIT Press, 1986.

Progressive Taxation and Equal Sacrifice

By H. PEYTON YOUNG*

Fairness is the dominant theme in almost every political debate about income tax policy.[1] Yet when it comes to actually assessing the treatment of different income groups, there is little or no agreement on how, or even whether, fairness can be meaningfully measured. The difficulty is that most criteria of vertical equity are based on the notion of equal sacrifice. While this idea was influential around the turn of the last century, it is now considered quite unfashionable, if not downright disreputable, since it relies heavily on interpersonal utility comparisons (Paul Samuelson, 1947; Richard Musgrave, 1959; Anthony Atkinson and Joseph Stiglitz, 1980). In spite of its dubious theoretical foundations, however, we propose to examine whether equal sacrifice may explain why observed tax rates have the particular structure that they do. In other words, is it a valid empirical principle?

Equal sacrifice is an elaboration of the notion that a rich person should pay more in taxes than a poor person because the former feels a given monetary loss to a lesser degree.[2] The case for it was put most succinctly by John Stuart Mill:

> As a government ought to make no distinction of persons or classes in the strength of their claims on it, whatever sacrifices it requires from them should be made to bear as nearly as possible with the same pressure upon all ...Equality of taxation, therefore, as a maxim of politics, means equality of sacrifice. [1848, p. 804]

This passage spawned a large and illustrious literature on sacrifice theory around the turn of the century (Henry Sidgwick, 1883; Arnold Jacob Cohen Stuart, 1889; Gustav Cassell, 1901; F. Y. Edgeworth, 1897, 1919; Arthur Pigou, 1928). Below we shall briefly review the various interpretations that have been given to the term "equal sacrifice." The point that bears emphasizing here is that Mill was suggesting the concept as a *political* principle. Equal sacrifice is a natural corollary of egalitarianism. If we consider Mill's statement in this light, then it is reasonable to ask whether equal sacrifice is discernible in the way that legislators actually do distribute the tax burden. Specifically, we shall ask whether different income groups give up approximately the same amount (alternatively, the same proportion) of their utility in paying taxes. The credibility of the answer will depend, of course, on whether the estimated form of the utility function accords well with estimates of utility derived from other sources, such as the finance literature. We shall find that it does.[3]

---

*H. Peyton Young is Professor of Public Policy, School of Public Affairs, University of Maryland, College Park, MD 20742. This work was completed while the author was a guest scholar in the Economic Studies Program at the Brookings Institution, Washington, DC 20036. The manuscript benefited from helpful comments by Marcus Berliant, Richard Musgrave, Joseph Pechman, Wolfram Richter, Michael Wallerstein, Stan Winer, and two anonymous referees. The author thanks Dean Foster and Eric Munz for assistance in analyzing the tax data. The research was supported by NSF grant no. SES-831-9530.

[1] For example, the title of the recent U.S. tax reform proposal was *Tax Reform for Fairness, Simplicity, and Economic Growth* (U.S. Department of the Treasury, 1984).

[2] Equal sacrifice does not necessarily imply (as some early authors erroneously assumed) that the rich should pay *proportionally* more of their incomes in tax (Samuelson, 1947, p. 247).

[3] An early attempt to estimate the marginal utility of income from tax data is due to Koichi Mera (1969). Irving Fisher (1927) suggested the reverse procedure: estimate the marginal utility of income from consumption data, and then substitute this into an equal sacrifice formula to determine the "just" rate of income tax progression. For related work, see Gabrielle Preinreich (1948), Otto Eckstein (1961), Robert H. Haveman (1965), and Burton Weisbrod (1968).

The data on which we test this hypothesis consist of federal income tax schedules in the United States over the period 1957–1987. During this period the income tax underwent a half-dozen substantial reforms.[4] The top bracket dropped from 91 percent in 1957 to 38.5 percent in 1987 to 33 percent today. In spite of these dramatic shifts, the distribution of the tax burden is explained quite well by the equal sacrifice model in most years. The post-1986 tax reform schedule is, however, a notable exception, as we shall presently see. In every case where a good fit is obtained, the estimated utility function exhibits constant proportional risk aversion with a coefficient between 1.5 and 1.7. These values are in good agreement with recent estimates based on cross-sectional studies of household demand for risky assets (Irwin Friend and Marshall E. Blume, 1975.) Similar results are obtained for recent nominal tax schedules in West Germany, Japan, and Italy.

The United Kingdom, like the United States, deviates significantly from the equal sacrifice model however. A possible explanation for this is that in both cases the schedules resulted from tax reforms in which great importance was attached to reducing the *number* of distinct tax brackets (so-called "tax simplification"). A small number of distinct brackets, with sizable jumps between the brackets, results in a choppy pattern for the average tax rate that does not fit the equal sacrifice model nearly as well as a gradually rising series of brackets.

Although we cannot draw definitive conclusions from such a limited set of data, the results suggest that equal sacrifice provides a reasonably accurate model of how the U.S. federal tax burden has been distributed among most taxpayers, at least until recently. It is also significant (or at least a remarkable coincidence) that the estimated elasticity of the utility of income is in agreement with estimates based on the demand for risky assets. At the lower and upper ends of the distribution, however, the equal sacrifice model does not fit the data well. One explanation is that at lower incomes the need to raise revenue forces an initial rate that is higher than equal sacrifice requires, whereas at higher incomes the need to preserve economic incentives holds the marginal rates below what equal sacrifice requires. These results appear to generalize to other industrialized countries, but more work on this aspect remains to be done.

I. Concepts of Equal Sacrifice

Any empirical test of the equal sacrifice hypothesis is complicated by the existence of several competing versions of the concept (Musgrave, 1959). The idea originally advanced by Mill was that everyone should suffer the same absolute loss of utility. That is, if $U(x)$ represents the utility corresponding to income level x, then the tax t as a function of x should satisfy

$$(1)\qquad U(x)-U(x-t)=s,$$

where s is the constant level of sacrifice for all income classes, x. This implies that the tax schedule takes the form

$$(2)\quad t=x-U^{-1}[U(x)-s]\quad \text{for all } x>0.$$

Ideally, individuals should be differentiated according to their particular utility functions. But this is impossible in practice, and, even if it were possible, would be based on false premises because it requires making fine-tuned interpersonal utility comparisons. A more plausible point of view is to consider $U(x)$ as a social norm—the utility function of a "representative" member of society (Lionel Robbins, 1938; Musgrave, 1959). In this sense no interpersonal utility comparisons are being made; rather, individuals are being treated as if they were all alike. This is a typical assumption in many types of economic models, including most treatments of optimal taxation.

Mill proposed using the Bernoullian utility function, which was the standard of his day. In this case a fixed percentage decrease in income represents the same loss of utility at every income level. Therefore, everyone sacrifices the same amount of utility if each person pays the same percent of income in tax.[5] It is noteworthy that this solution, which is considered by many to be the simplest and fairest, can be justified on equal sacrifice grounds.

---

[4] In each of the periods 1954–1963, 1964, 1965–1976, 1977–1978, 1979–1981, U.S. individual income tax schedules remained fairly stable. From 1982–1987 major tax reforms were instituted through a series of transition schedules.

Subsequent to Mill, the equal sacrifice doctrine was elaborated in several directions. Cohen Stuart (1889) proposed that everyone should suffer the same *relative* loss in utility. If r is the rate of loss in utility, then for all $x > 0$,

$$U(x-t)/U(x) = 1 - r. \tag{3}$$

This criterion is known as "equal proportionate sacrifice."

For present purposes there is no need to distinguish the case of equal absolute from equal proportionate sacrifice, because equal proportionate sacrifice is nothing but equal absolute sacrifice relative to a different utility function. Namely, if we take the logarithm of both sides of (3), then we see that equal proportional sacrifice with respect to $U(x)$ amounts to equal absolute sacrifice with respect to $\ln U(x)$.[6]

II. A Test for Equal Sacrifice

To make any progress on testing equal sacrifice, it would appear that we must first specify the form of the utility function. Actually, this is not so. Instead, we shall postulate that equal sacrifice holds for some (unknown) utility function, and then show that important information about the utility function can be derived directly from the tax data. The equal sacrifice hypothesis will be plausible if: (i) the estimated utility function is reasonably consistent with utility theory; and (ii) the equal sacrifice schedule derived from this utility function fits the empirical tax data.

In the modern theory of risk bearing, two parameters play a key role in defining the utility function: the coefficient of absolute risk aversion $R(x) = -U''(x)/U'(x)$ and the coefficient of proportional risk aversion $C(x) = -xU''(x)/U'(x)$ (John Pratt, 1964, Kenneth J. Arrow, 1971.) It is now generally accepted that the coefficient of absolute risk aversion is decreasing, while the coefficient of proportional risk aversion is more or less constant. Constant proportional risk aversion implies that people hold a constant *proportion* of their wealth in any one class of risky assets as their wealth varies. Empirical studies of household wealth have tended to support this hypothesis, and the coefficient C has been estimated to be greater than 1, and probably in the general neighborhood of 2 (Friend and Blume, 1975). This implies that the utility function is of the form:

$$U(x) = -A(x)^{1-C} + B, \qquad A > 0, \quad C > 1. \tag{4}$$

Turning now to the test for equal sacrifice, the first step is to examine the behavior of C as a function of x. It turns out that C can be estimated directly from the tax data. To see this, consider an empirically given tax schedule $t = f(x)$, where t is the amount of tax paid by persons at income level x. Assume that there exists a utility function $U(x)$ such that the loss of utility at all levels of income x is approximately constant:

$$U(x) - U(x-t) = s.$$

Dividing both sides by t, we obtain

$$[U(x) - U(x-t)]/t = s/t. \tag{5}$$

By the Mean Value Theorem, the left-hand side of (5) is equal to the derivative of U at some intermediate value, w, between x and

---

[5] Mill defined taxable income to be income net of subsistence requirements as well as savings.

[6] A third variation of the equal sacrifice theme is to minimize aggregate sacrifice. This means that taxes should be distributed so as to minimize the total loss of utility summed over all individuals. The solution (assuming that the utility of income is increasing and strictly concave) is to equalize everyone's after-tax income (Edgeworth, 1897). This welfare maximization approach can be made much more appealing by employing a more realistic utility function—one that incorporates, for example, the tradeoff between income and leisure (James A. Mirrlees, 1971; J. K. Seade, 1977).

$x-t$. Of course, w cannot be known precisely unless we know $U(x)$, which is what we are trying to estimate. Nevertheless, it turns out that w can be estimated quite accurately without knowing U. To see this, assume that the coefficient C is more or less constant in some neighborhood that includes x and $x-t$. Then, in this neighborhood, $U(x)=-Ax^{1-C}+B$, and without loss of generality we may take $A=1$ and $B=0$. Thus $U'(x)=(C-1)x^{-C}$ and the defining equation for w is

$$\begin{aligned} U'(w) &= (C-1)w^{-C} \\ &= [U(x)-U(x-t)]/t, \\ &= \left[(x-t)^{1-C}-x^{1-C}\right]/t. \end{aligned}$$

After some algebraic manipulation, we find that

$$w/x=\left(\frac{(C-1)(t/x)}{(1-t/x)^{1-C}-1}\right)^{1/C}.$$

Take a typical value of t/x, say $t/x=0.2$, and substitute in various plausible values for C. As the following table shows, the resulting value of w/x (and hence of w, given a specific value of x) is quite insensitive to the choice of C.

C	w/x
3	0.893
2.5	0.894
2.0	0.894
1.5	0.895
1.1	0.896

The upshot is that we may safely choose any value of C in this range in order to estimate w. The value $C=2$ seems like a good choice and leads to the particularly simple formula $w=\sqrt{x(x-t)}$. From this and equation (5) we therefore have

$$U'\left(\sqrt{x(x-t)}\right)=s/t.$$

Without loss of generality we may take $s=1$. Taking logarithms we obtain

$$\ln U'\left(\sqrt{x(x-t)}\right)=-\ln t. \tag{6}$$

Recall now that we are attempting to estimate the coefficient of proportional risk aversion $-zU''(z)/U'(z)$. This is the rate of change of $-\ln U'(z)$ with respect to $\ln z$, which is $d(-\ln U'(z)/d(\ln z)=-[U''(z)/U'(z)]/[1/z]$. Let $X=\ln z$ and let $Y=-\ln U'(z)$. If we regress Y against X, then the slope of the regression line will be an estimate of C.

Let $z=\sqrt{x(x-t)}$. Then

$$X=\ln z=\ln\sqrt{x(x-t)}$$

and, by (6), $Y=-\ln U'(\sqrt{x(x-t)})=\ln t$. Thus, we wish to regress $Y=\ln t$ against $X=\ln\sqrt{x(x-t)}=(1/2)\ln x(x-t)$ for various levels of tax t and pre-tax income x. The higher the R^2 is, the more plausible is the hypothesis that C is independent of x, and that the tax equalizes sacrifice relative to an isoelastic utility function.

III. Empirical Results

We illustrate the approach for United States tax data in 1957. Table 1 shows tax paid as a function of Adjusted Gross Income, which is the closest approximation we have to the effective tax schedule.[7]

The first step is to estimate the coefficient of proportional risk aversion as described in the preceding section. Let $Y=\ln t$ and $X=(1/2)\ln x(x-t)$ for the various values of x and t in the table. If the equal sacrifice hypothesis is correct relative to an isoelastic utility function, then we would expect to see an approximately linear relationship between X and Y, and the slope of the regression line will be an estimate of the coefficient C. Figure 1a shows that this hypothesis is

---

[7] It would be far better, of course, to make this estimation relative to total personal income rather than Adjusted Gross Income. Unfortunately, these data are not available.

TABLE 1—FEDERAL TAX PAID AS A FUNCTION OF ADJUSTED GROSS INCOME (AGI)
UNITED STATES, 1957[a]

Income Class	Percent of Total Returns in the Class	Average Income	Tax Paid	Percent
Under $600	6.5	328	0	0
600– 1,000	5.0	798	13	1.6
1,000– 1,500	7.0	1,241	48	3.9
1,500– 2,000	6.2	1,752	90	5.1
2,000– 2,500	6.5	2,252	136	6.0
2,500– 3,000	6.4	2,748	188	6.8
3,000– 3,500	6.5	3,246	247	7.6
3,500– 4,000	6.6	3,751	309	8.2
4,000– 4,500	6.7	4,250	370	8.7
4,500– 5,000	6.5	4,748	431	9.1
5,000– 6,000	11.0	5,474	525	9.6
6,000– 7,000	7.9	6,472	690	10.7
7,000– 8,000	5.4	7,466	870	11.7
8,000– 9,000	3.5	8,467	1,065	12.6
9,000– 10,000	2.2	9,458	1,257	13.3
10,000– 15,000	3.7	11,744	1,740	14.8
15,000– 20,000	0.9	17,112	3,013	17.6
20,000– 25,000	0.4	22,256	4,468	20.1
25,000– 50,000	0.6	33,373	8,472	25.4
50,000–1000,000	0.2	65,652	23,262	35.4
Above 100,000	0.3	–	–	–

[a]*Source:* Statistics of Income, U.S. Internal Revenue Service, 1957.

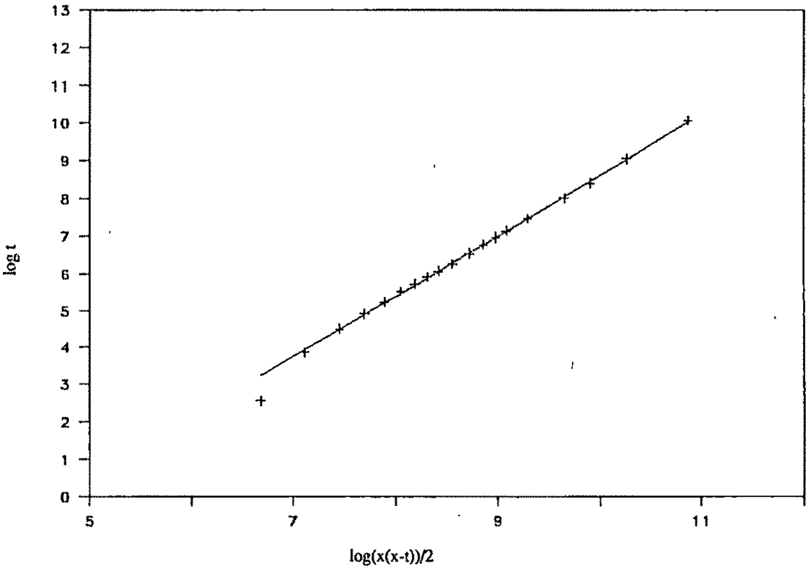


FIGURE 1a. SLOPE ESTIMATE OF C, U.S. EFFECTIVE TAX SCHEDULE, 1957, $C = 1.61$, S.E. $= 0.008$, $R^2 = 99.9$

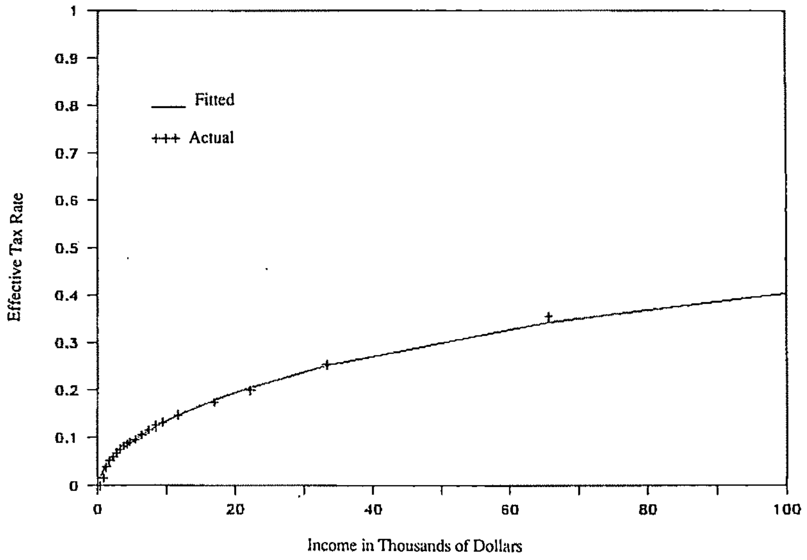


FIGURE 1b. EQUAL SACRIFICE TAX FITTED TO U.S. EFFECTIVE SCHEDULE, 1957. ES TAX $= x-(x^{-0.610}+0.000337)^{-1/0.610}$. INCOME RANGE: $\$1,000 \leq x \leq \$100,000$. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 3.8 PERCENT

strongly confirmed for adjusted gross incomes above \$1,000, which represents 88.2 percent of all returns in 1957.[8] The estimated value of C is 1.61, the standard error of the estimate is 0.008, and the R^2 is 99.9 percent. This finding does not confirm the hypothesis of equal sacrifice itself; it merely gives us confidence in the estimated value of C and the isoelasticity of the utility function *assuming* that equal sacrifice holds. To a good approximation, therefore, the utility function may be written as $U(x)=-x^{-0.61}$.

Next, we plot the differences $U(x)-U(x-t)$ to estimate the level of sacrifice, s. The value of s has no absolute significance, of course, since it depends on the scaling of the utility function. Nevertheless, it is a necessary parameter for computing the equal sacrifice tax, given that the utility function has been specified. Again, treating incomes below \$1,000 as outliers, the estimated mean level of sacrifice is $s=3.37\times10^{-4}$ and the standard deviation is 0.1×10^{-4}, which is about 3 percent of the mean.

The final step is to use the estimated values of C and of s to compute a fitted tax schedule $t=x-(x^{1-C}+s)^{1/(1-C)}$. This is shown in Figure 1b. For incomes between \$1,000 and \$100,000, the ratio of the equal sacrifice tax to the actual tax has a coefficient of variation of ±3.8 percent. To a very good approximation, therefore, the effective tax rate in 1957 is consistent with equal absolute sacrifice relative to the isoelastic utility function $U(x)=-x^{-0.61}$.[9]

---

[8] The data point corresponding to the income class \$600–\$1,000 is treated as an outlier (see Figure 1a). If this point is included in the estimation, then the estimated value of C is 1.67 and the standard error is 0.033. As noted earlier, there are good reasons to expect a departure from the equal sacrifice model at the lower end of the income distribution. Hence we have estimated C after excluding the lower tail, where linearity does not appear to be confirmed.

[9] It should be noted that these results are also consistent with equal *proportionate* sacrifice relative to the utility function $U(x)=\exp[-Ax^{-0.61}+B]$. While this

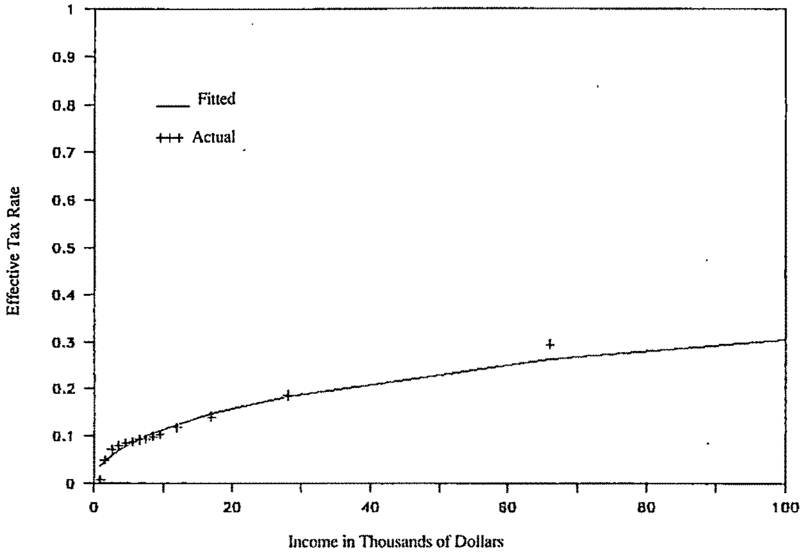


FIGURE 2. EQUAL SACRIFICE TAX FITTED TO U.S. EFFECTIVE SCHEDULE, 1967. ES TAX = $(x^{-0.519}+0.000566)^{-1/0.519}$. INCOME RANGE: $\$3{,}000 \leq x \leq \$100{,}000$. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 7.2 PERCENT

Similar results are obtained for the tax years 1967 and 1977, though the fit is somewhat less good. (See Figures 2 and 3). The estimation for 1987 could not be done because the relevant data are not yet available from the Internal Revenue Service. The nominal tax schedules are available, however, and this is the estimation we turn to next.

It is quite conceivable that the schedule of published rates (the nominal schedule) is also consistent with equal sacrifice. The hypothesis here is that the public is at least as sensitive to the apparent distribution of tax rates as they are to the effective rates. In other words, the appearance of equity may be as important as equity in fact. Certainly, the question seems worthy of investigation. Furthermore, the nominal schedules have a distinct advantage over the effective schedules in that they are not subject to measurement error.

For purposes of this analysis we chose U.S. Schedule X, which applies to individuals. A good fit is obtained for the years 1957, 1967, and 1977, except at the very lower and upper ends of the income scale (Figure 4 is illustrative.)[10] The estimated elasticities are somewhat higher than for the corresponding effective schedules, as is to be expected, since the higher the elasticity of marginal utility, the greater the progressivity of the schedule (see Table 2).

---

is a less standard representation of utility than the isoelastic functions, it is not wholly unreasonable. In fact, the coefficient of proportional risk aversion for $U(x)=\exp[-Ax^{-p}+B]$ is $1+p-Ap/x^{p}$, which is slightly increasing but very close to the coefficient for $-Ax^{-p}+B$ when x is large relative to A. In other words, the two functions are scarcely distinguishable with respect to risk aversion, and hence the model cannot really say whether equal absolute or equal proportionate sacrifice is more credible. Given the general preference for the isoelastic representation of utility in the literature, we shall interpret our results as implying equal absolute sacrifice relative to an isoelastic utility function.

[10] The coefficient of variation for the equal sacrifice tax divided by the actual tax is 8.8 percent for 1957, 5.6 percent for 1967, and 6.9 percent for 1977.

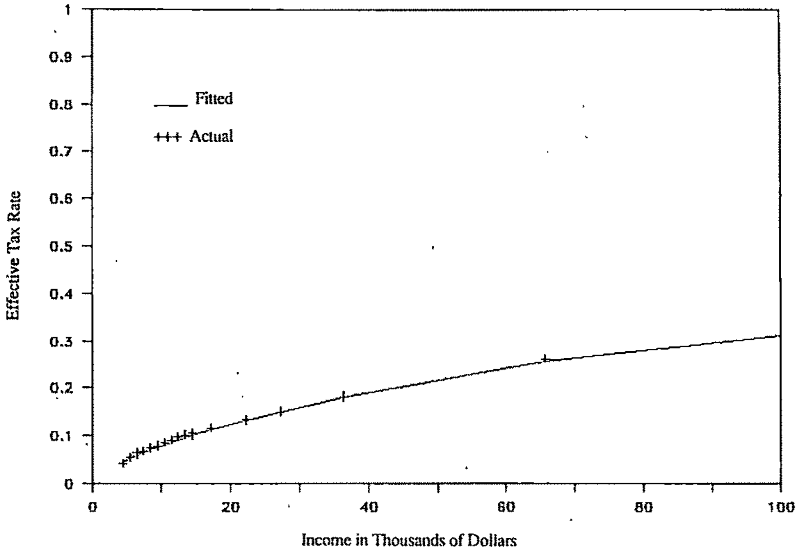


FIGURE 3. EQUAL SACRIFICE TAX FITTED TO U.S. EFFECTIVE SCHEDULE, 1977. ES TAX $= x-(x^{-0.718}+0.00008)^{-1/0.718}$. INCOME RANGE: $\$4,000 \leq x \leq \$100,000$. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 4.0 PERCENT

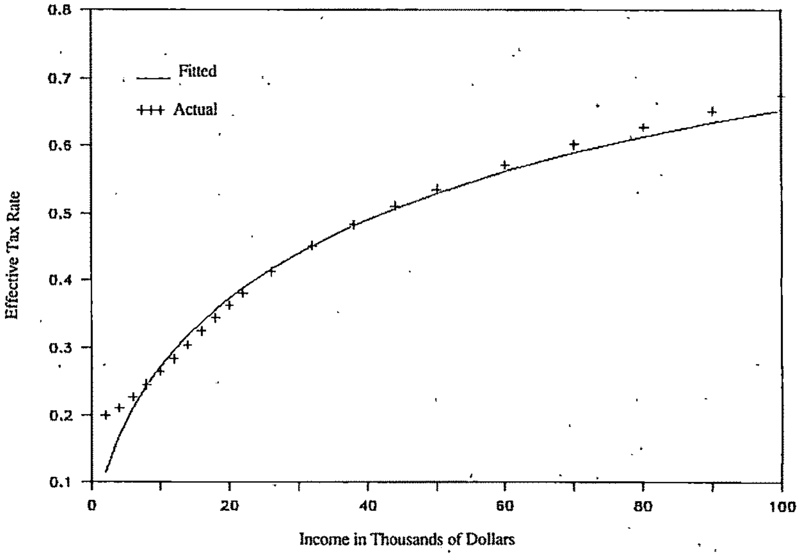


FIGURE 4. EQUAL SACRIFICE TAX FITTED TO U.S. NOMINAL SCHEDULE, 1957. ES TAX $= x-(x^{-0.631}+0.000664)^{-1/0.631}$. INCOME RANGE: $\$3,000 \leq x \leq \$100,000$. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 5.2 PERCENT

TABLE 2—ESTIMATED ELASTICITY OF MARGINAL UTILITY, U.S. NOMINAL AND EFFECTIVE TAX SCHEDULES, 1957–1987.

	Nominal	Effective
1957	1.63	1.61
1967	1.53	1.52
1977	1.79	1.72
1987	1.37	n/a

By contrast, the 1987 nominal schedule does not fit the equal sacrifice model very well (see Figure 5). The reason is that the tax is nearly flat-rate for incomes up to \$16,800 and hence progressivity is almost nil in this range.[11] Above \$16,800 progressivity is modest but fairly steady as marginal rates rise from 15 percent to 38.5 percent. This example clearly demonstrates that the equal sacrifice model does not explain all tax schedules. But it also shows that the model is not tautological: There exist perfectly reasonable tax schedules that do *not* support an equal sacrifice interpretation, at least not relative to an isoelastic utility function.

A possible explanation for the departure from equal sacrifice in 1987 (as compared with prior years) is the emphasis placed in the 1986 Tax Reform Act on "simplifying" the tax structure. One of the supposed simplifications was to reduce the number of brackets. But no schedule composed of just two or three marginal rates will fit the equal sacrifice model well, because equal sacrifice relative to any smooth utility function implies a continuously varying marginal rate.[12] The converse is not true: just because a tax schedule exhibits a continuously increasing marginal rate, does not imply that it is "almost" an equal sacrifice tax. Indeed, an equal sacrifice tax relative to an isoelastic utility function (with $1 < C < 2$) has a very special shape: the effective tax rate t/x is concave, increases continuously from zero and is asymptotic to 100 percent as income goes to infinity. Actually, more is true: once two points on the schedule are chosen—that is, once the tax is specified for two distinct incomes—then the equal sacrifice schedule and the corresponding utility function are fully determined. So, it would be highly coincidental if an arbitrarily chosen tax rate schedule (even one with many brackets) were to meet these requirements. For example, a tax rate schedule in which the effective rate t/x is first concave, then convex, then concave (as in the 1987 U.S. Schedule) does not fit the equal sacrifice model that we have described.

While earlier U.S. tax schedules are generally consistent with the equal sacrifice hypothesis over most of the income distribution, they do not fit the model at the lower end. The reason is that the tax brackets are not graduated finely enough for low incomes. Indeed, the equal sacrifice model (with $C > 1$) requires that the marginal tax rate decrease continuously to zero as income approaches zero. Any schedule based on a finite number of brackets obviously violates this condition. Given that the initial brackets in the U.S. schedules varied between 11 percent and 20 percent during the period 1957–1987, it can hardly be expected that the fit would be good at the lower end of the income scale. The reason for such large initial rates is a matter of fiscal arithmetic: In order to capture enough revenue, one must tax where the income is, and the lion's share of taxable income lies in the income brackets just above the minimum exemption level. So it is almost necessary for the initial marginal rate to be large, or at least to rise very steeply. This constraint may override considerations of fairness at the bottom of the income scale.

At the upper end of the income distribution we find departures from the equal sacrifice model for quite a different reason. Marginal tax rates must be truncated well below 100 percent in order to provide adequate incentives for people to work and invest. This is inconsistent with equal sacrifice

---

[11] Income up to \$1,800 is taxed at 11 percent, then at 15 percent up to a total of \$16,800. The fit is even worse for 1988, where the initial rate is 15 percent on the first \$17,850, 28 percent up to \$43,150, and 33 percent up to \$89,560 (Internal Revenue Service, 1987, 1988.)

[12] From $U(x) - U(x-t) = s$ it follows by differentiation that $dt/dx = 1 - U'(x)/U'(x-t)$. Hence, if U' is continuous, then so is the marginal rate dt/dx.

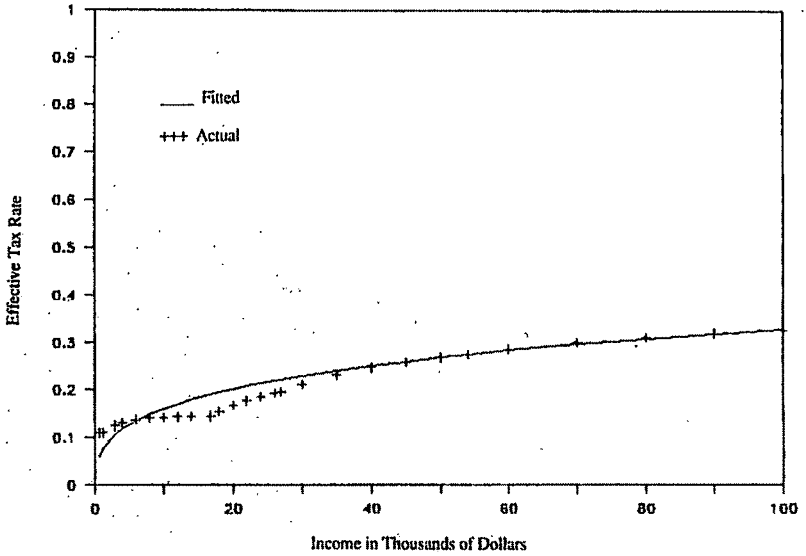


FIGURE 5. EQUAL SACRIFICE TAX FITTED TO U.S. NOMINAL SCHEDULE, 1987. ES TAX $= x-(x^{-0.373}+0.00218)^{-1/0.373}$. INCOME RANGE: $\$3{,}000 \le x \le \$100{,}000$. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 10.3 PERCENT

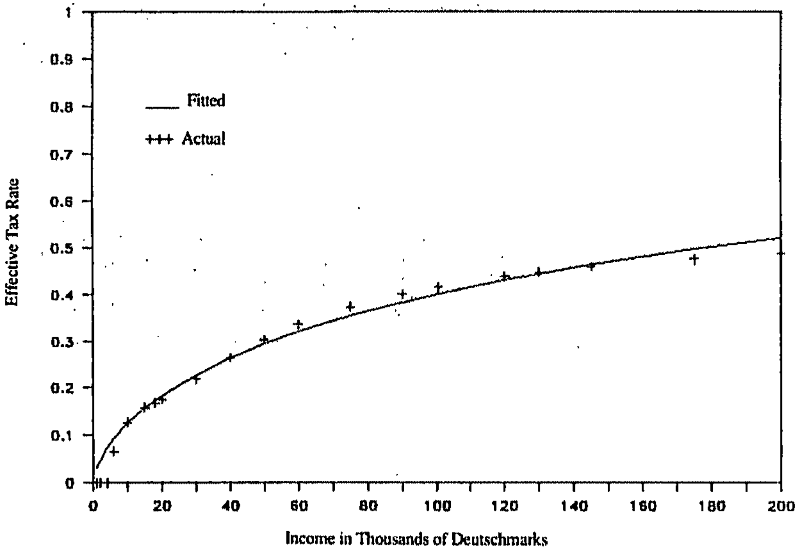


FIGURE 6. EQUAL SACRIFICE TAX FITTED TO WEST GERMAN NOMINAL SCHEDULE, 1984. ES TAX $= x-(x^{-0.633}+0.000260)^{-1/0.633}$. INCOME RANGE: DM $10{,}000 \le x \le$ DM $200{,}000$. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 3.2 PERCENT

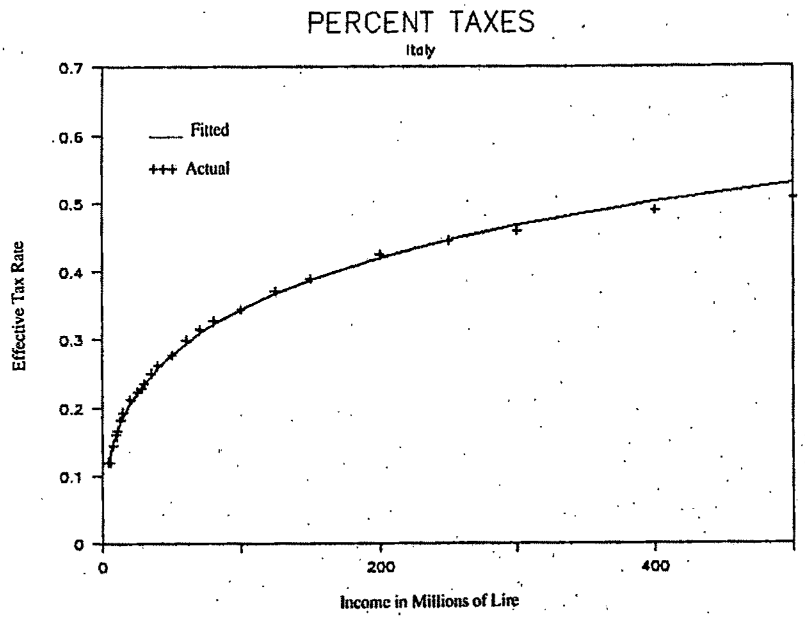


FIGURE 7. EQUAL SACRIFICE TAX FITTED TO ITALIAN NOMINAL SCHEDULE, 1987. ES TAX $= x-(x^{-0.403}+0.00179)^{-1/0.403}$. INCOME RANGE: $4 \leq x \leq 500$ MILLION LIRE. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 3.9 PERCENT

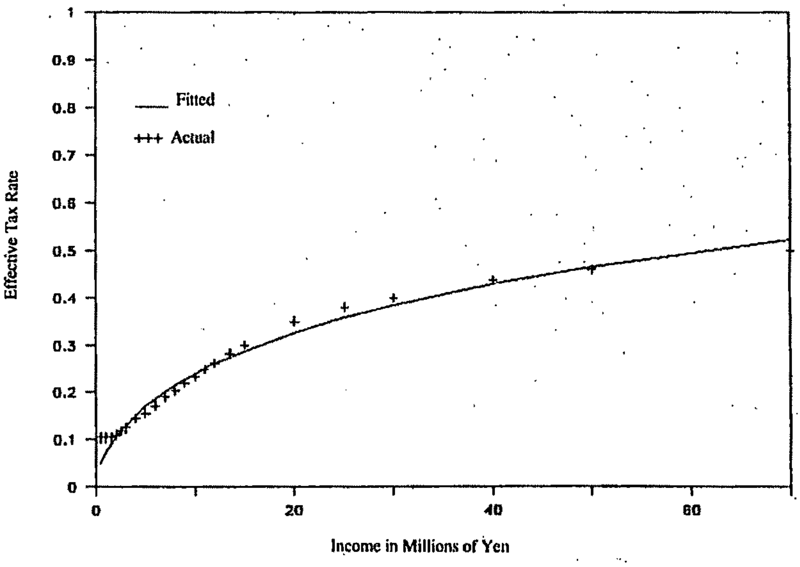


FIGURE 8. EQUAL SACRIFICE TAX FITTED TO JAPANESE NOMINAL SCHEDULE, 1987. ES TAX $= x-(x^{-0.587}+0.0448)^{-1/0.587}$. INCOME RANGE: $1.5 \leq x \leq 70$ MILLION YEN. COEFFICIENT OF VARIATION OF ES TAX/ACTUAL TAX = 5.4 PERCENT

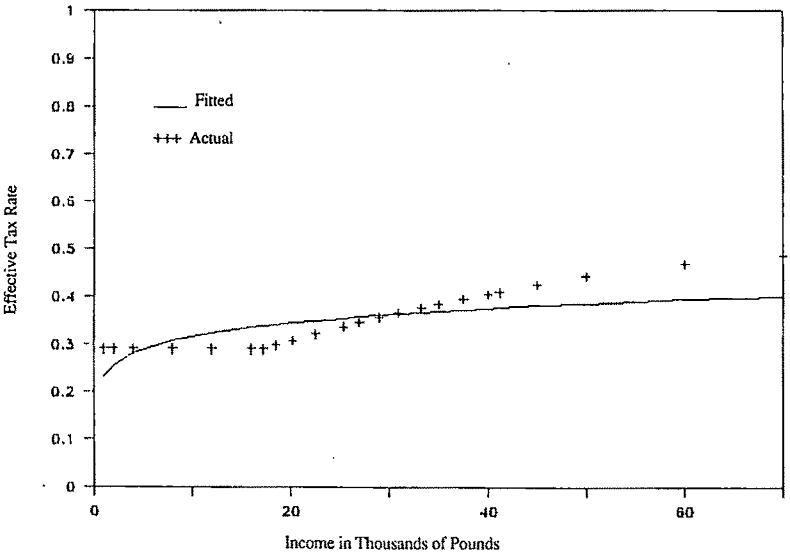


FIGURE 9. EQUAL SACRIFICE TAX FITTED TO U.K. NOMINAL SCHEDULE, 1987. ES TAX $= x - (x^{-0.163} + 0.0143)^{-1/0.163}$. INCOME RANGE: $1 \leq x \leq 70$ THOUSAND POUNDS. COEFFICIENT OF VARIATION OF ES TAX = 10.4 PERCENT

relative to an isoelastic utility function, however, which requires that tax rates gradually approach 100 percent. For high incomes, therefore, the departure from equal sacrifice may be due to efficiency considerations, whereas for low incomes it is probably due to revenue requirements. The middle- to upper-middle income range is where considerations of vertical equity can be given somewhat freer rein.

IV. Data from Other Countries

It is natural to ask whether the preceding results are in some way peculiar to the United States. To investigate this possibility in a preliminary way, we chose four major industrialized countries—West Germany, Italy, Japan, and the United Kingdom—and analyzed the most recent nominal tax schedules available to us. The results are illustrated in Figures 6–9. The equal sacrifice model gives an excellent fit for West Germany, Italy, and Japan, and the estimated coefficients are 1.63, 1.40, and 1.59, respectively. It is rather remarkable that these values all lie within such a narrow range, and that they are so similar to the U.S. results. Italy is particularly noteworthy because the nominal schedule fits the equal sacrifice model very closely even at the low end of the income scale. All three countries exhibit a much more finely graduated rate structure than the current U.S. schedule does. The United Kingdom, however, is similar to the United States in that the 1987 schedule does not fit the equal sacrifice model at all well. The reason is that it employs a flat-rate tax on taxable income up to £17,200. and a mildly progressive series of rates thereafter. As in the case of the recent U.S. tax reform, this appears to be the result of a political compromise in which the drive toward a flat-rate tax had to be modified by demands for progressive treatment of the well-to-do.

V. Conclusion

In this paper we have described a general method for testing whether a tax schedule exhibits equal sacrifice relative to an isoelastic utility function. The technique can be applied to any schedule, whether nominal or effective, and even to particular portions of a given schedule. The method estimates the coefficient of risk aversion in the utility function, the level of sacrifice at each level of income, and the tax function that would be implied by equal sacrifice. The latter may then be compared with the actual schedule to see how far the actual tax deviates from equal sacrifice on different portions of the income distribution.

Over much of the postwar period, the observed tax rates in the United States conformed to the equal sacrifice model quite closely, and the estimated curvature of the utility function was fairly stable. Of course, obtaining a reasonably good fit does not prove causation. Nor do we have any direct evidence that legislators actually invoked equal sacrifice arguments in proposing the rate structures that we observe. It seems likely, however, that intuitive notions of "relative sacrifice" and "ability to pay" are one factor in the way that legislators evaluate the fairness of tax proposals. And it does not seem too far-fetched to suppose that the aggregate of these intuitions, as expressed in a majority vote, might come close to an equal sacrifice tax relative to an "average" utility function.

Several questions remain to be explored. First, it would be interesting to know whether these results hold up when we analyze the effective rather than the nominal schedules in other industrialized countries. Even for the United States, it would be far better to carry out the analysis relative to full income, rather than Adjusted Gross Income, as we were forced to do because of lack of data.

Second, it may well be that some other theory can explain why tax schedules merely look like equal sacrifice schedules. Such a theory would need to explain why effective rates t/x tend to be concave, and more specifically why they tend to fit functions of the form

$$t = x - [x^{-p} + s]^{-1/p}, \qquad s > 0, \quad 0 < p < 1. \tag{7}$$

Third, it may be that equal sacrifice has some other explanation or justification than the traditional utilitarian one. One idea along these lines is the following. If we hypothesize that taxes are distributed according to *some* measure of ability to pay (not necessarily a utilitarian one), and if we suppose further that the criterion applies not only to the distribution of the whole tax, but to every tax increase (or decrease), then the measure of ability to pay must be equal sacrifice relative to a social utility of income (Young, 1988).[13] This argument suggests that the equal sacrifice "look" might result from legislators trying to balance equity in incremental changes that they make to the tax distribution with equity in the overall result.

Resolving these issues is beyond the scope of the present paper. The evidence suggests, however, that equal sacrifice may play a significant role in the way that people think about taxation, and that it needs to be taken more seriously as the 'maxim of politics' that Mill claimed it to be.

---

[13] We assume here that taxes are positive, continuous, strictly increasing in income, strictly increasing as a function of the total tax burden, and that marginal rates are less than 100 percent.

REFERENCES

Arrow, Kenneth J., *Essays in the Theory of Risk Bearing*, Amsterdam: North-Holland, 1971.

Atkinson, Anthony and Stiglitz, Joseph, *Lectures on Public Economics*, New York: McGraw-Hill, 1980.

Cassell, Gustav, "The Theory of Progressive Taxation," *Economic Journal*, December 1901, *11*, 481–91.

Stuart, Cohen and Cohen, Arnold Jacob, 1889, "On Progressive Taxation," in R. A. Musgrave and A. T. Peacock, eds., "*Classics in the Theory of Public Finance*," New York: MacMillan, 1958.

Eckstein, Otto, "A Survey of the Theory of Public Expenditure Criteria," National Bureau of Economic Research, *Public Finances: Needs, Sources, and Utilization*, Princeton: Princeton University Press, 1961.

Edgeworth, F. Y., "The Pure Theory of Taxation," in R. A. Musgrave and A. T. Peacock, eds., *Classics in the Theory of Public Finance*, New York: MacMillan, 1958.

______, "Methods of Graduating Taxes on Income and Capital," *Economic Journal*, 1919, *29*, 138–53.

Fisher, Irving, "A Statistical Method for Measuring Marginal Utility and Testing the Justice of a Progressive Income Tax," *Economic Essays in Honor of John Bates Clark*, New York: MacMillan, 1927.

Friend, Irving and Blume, Marshall E., "The Demand for Risky Assets," *American Economic Review*, December 1975, *65*, 900–22.

Haveman, Robert H., *Water Resource Investment and the Public Interest*, Nashville: Vanderbilt University Press, 1965.

Ishi, Hiromitsu, *The Japanese Tax System*, Oxford: Clarendon Press, 1989.

Mera, Koichi, "Expermental Determination of Relative Marginal Utilities," *Quarterly Journal of Economics*, August 1969, *83*, 464–77.

Mill, John Stuart, *Principles of Political Economy* (1848), London: Longmans Green, 1917.

Mirrlees, James A., "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, April 1971, *38*, 175–208.

Musgrave, Richard, *The Theory of Public Finance*, New York: McGraw-Hill, 1959.

Pfingsten, Andreas, *The Measurement of Tax Progression*, unpublished doctoral dissertation, Karlsruhe, West Germany: University of Karlsruhe, 1985.

Pigou, Arthur, *A Study in Public Finance*, London: MacMillan, 1928.

Pratt, John, "Risk Aversion in the Small and in the Large," *Econometrica*, January-April 1964, *32*, 122–36.

Preinreich, Gabrielle, "Progressive Taxation and Proportionate Sacrifice," *American Economic Review*, March 1948, *38*, 103–17.

Price Waterhouse, *Individual Taxes: A Worldwide Summary*, New York: Price Waterhouse, 1987.

Robbins, Lionel, "Interpersonal Comparisons of Utility," *Economic Journal*, December 1938, *48*, 635–41.

Samuelson, Paul, *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

Seade, J. K., "On the Shape of Optimal Tax Schedules," *Journal of Public Economics*, April 1977, *7*, 203–35.

Sidgwick, Henry, *The Principles of Political Economy*, London: MacMillan, 1983.

Weisbrod, Burton, "Income Redistribution Effects and Benefit-Cost Analysis," in: S. B. Chase ed., *Problems in Public Expenditure Analysis*, Washington: The Brookings Institution, 1968.

Young, H. Peyton, "Distributive Justice in Taxation," *Journal of Economic Theory*, April 1988, *44*, 321–35.

U.S. Treasury Department, *Statistics of Income: Individual Tax Returns*, 1957–1977.

______, *Tax Reform for Fairness, Simplicity, and Economic Growth*. Washington, DC: USGPO, 1984.

______, Internal Revenue Service, Form 1040, Schedule X, Washington, DC: USGPO, 1987, 1988.

Job Accessibility and Racial Differences in Youth Employment Rates

By KEITH R. IHLANFELDT AND DAVID L. SJOQUIST*

Over the past thirty years the secular trend in black youth employment rates has been downward, both absolutely and relative to the employment rates for white youth. Currently, the employment rate for black teenagers is only about one-half as high as the employment rate for white teenagers (0.49 for whites and 0.27 for blacks in November, 1988).[1] Despite considerable research effort, much of the trend as well as the existing gap in employment among black and white youth remains unexplained, which obviously makes it difficult to formulate policies that would effectively mitigate this important social problem.[2]

One hypothesis that may help explain both the trend and the racial gap in youth employment is that the suburbanization of low-skill jobs and continued housing market segregation have acted together to reduce the job opportunities of black youth in comparison to those available to white youth. This is an attractive hypothesis since evidence suggests that inner-city job access has declined over time for low-skilled workers (see, for example, John Kasarda, 1985). While the origins of this hypothesis can be traced back to John Kain (1968), who argued that housing segregation reduces the employment opportunities of all blacks, in recent years it has been advanced by both prominent economists (Robert Gordon, 1982; Michael Borus, 1982) and sociologists (Kasarda, 1985; William Wilson, 1987) as one of the probable keys to the resolution of the black youth employment enigma.

The first test of the "job access" hypothesis as it relates to youth was conducted by David Ellwood (1986) using data from the Chicago metropolitan area. His results suggested that young workers are sufficiently "fluid" in their commuting patterns to overcome any problems arising from an absence of nearby jobs. He therefore concluded, "Thus the problem isn't space. It's race." (p. 181) This same conclusion was reached by Jonathan Leonard (1986a), who conducted a study similar to Ellwood's using data for the Los Angeles metropolitan area. In both stud-

---

*Policy Research Program and Department of Economics, Georgia State University, Atlanta, GA 30303. We wish to thank the anonymous referee for his valuable comments and Joon Song, Laura Heuchen, and Bob Eisenstadt for their able assistance. We benefited from discussions with John Jackson, Lou Ederington, Marty Fraiser, and Jane Leonard. Financial support for this project was provided by the Economic Development Administration of the U.S. Department of Commerce and the Research Program Committee, College of Business Administration, Georgia State University. The statements, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the Economic Development Administration.

[1] The employment rates for whites (blacks) in November, 1988 were 0.42 (0.21) and 0.68 (0.45) for 16–19-year-olds enrolled and not enrolled in school, respectively, and 0.58 (0.43) and 0.79 (0.66) for 20–24-year-olds enrolled and not enrolled, respectively, U.S. Department of Labor (1988).

[2] Part of the declining trend in the relative employment rate of black teenagers can be attributed to the migration of southern blacks from rural locations, where youth were in high demand as farm laborers, to urban areas (John Cogan, 1982). However, why these new locations offered a dearth of job opportunities for blacks is not clearly understood. Furthermore, since these population movements have long since subsided, they cannot explain the continued deterioration of the labor market position of black youth that occurred after 1970. Regarding the current differential in black/white employment rates, recent studies (Martin Feldstein and David Ellwood, 1982; Linda Leighton and Jacob Mincer, 1982) have found that less than half is explicable by racial differences in personal and family variables.

ies youth employment rates measured at the census tract level were regressed on variables describing job proximity and the socioeconomic characteristics of the tract's residents. While the job access variables were generally significant in both studies, estimated coefficients were uniformly small in magnitude. However, Leonard (1986b) has suggested Ellwood's measures of job accessibility may have been unreliable since their construction was based on small samples. Leonard's own measure of job access was the number of blue-collar jobs within a 15-minute commute of each census tract divided by the population 16 years of age and older of the commuting zone. Measurement error likely also affects this variable since the overwhelming majority of youth do not work in blue-collar jobs and the population of the commuting zone may poorly represent the number of workers competing for jobs typically held by youth.

In this paper the relationship between the nearness of jobs and youth job probability is explored using measures of job access that do not suffer from the limitations of those employed in previous work. Other improvements of the present study include the use of a richer set of control variables and the estimation of separate equations for black and white youth broken down by age and enrollment status, and for different metropolitan areas.

The data, the estimating equation, and the results obtained for black and white youth living in the Philadelphia metropolitan area are presented in Section I. The nearness to jobs was found to have a strong effect on the job probability of both black and white youth, regardless of their age, enrollment status, or whether they live at home or on their own. Estimates of the role that differential job accessibility plays in explaining the racial gap in youth employment rates indicate that from 33 percent to 54 percent of the gap can be attributed to space, depending on the group considered. Results for Chicago and Los Angeles are presented in Section II. These results lend additional support to our conclusion that poor job access is a significant contributor to the joblessness of black youth.

I. The Estimating Equation and Results for Philadelphia Youth

The location of jobs vis-à-vis the youth's residence can impinge on his (her) job probability for two distinct reasons. First, assuming the youth possesses full information of the spatial distribution of job openings and the wage gradient is relatively flat, for example the wage is fixed at the minimum wage, then as the required commuting distance increases, the wage net of travel expense declines, which decreases the likelihood that the net wage will exceed the youth's reservation wage. In comparison to higher wage workers, this effect may be particularly strong for the typical youth, since for any given distance, travel costs are a higher percentage of his earnings and his travel time is greater because he more frequently must rely on slower modes of transportation, for example, walking, bicycling, or buses. Second, as documented by Harry Holzer (1987), for both white and black youths, the most frequently used methods of job search are checking with friends and relatives and direct applications without referrals. The reliance on these informal methods of search suggest that the youth's information on available job opportunities may decay rapidly with distance from home.

Conceptually, the ideal measure of job accessibility would be the minimum distance the individual marginal youth would be required to commute if a job were taken. As a measure of this, we employed the mean travel time of workers living within the youth's community. As Ellwood notes, travel time is particularly appealing as a job access measure because it reflects actual worker behavior. If jobs are nearby, commuting time will be low. Conversely, if jobs can't be found nearby, travel time will be high.

The primary data utilized to investigate the relationship between mean travel time and youth job probability were for the Philadelphia SMSA and came from the Public Use Sample (PUS) of the 1980 U.S. Census of Population. Philadelphia offered a number of advantages that permitted the strongest possible tests that could be made of the job access hypothesis with PUS data.

First, greater subcounty spatial identification is provided for Philadelphia than for any other SMSA identified on the *A* (5 percent) Sample of the PUS.[3] Twenty-six areas, representing planning districts, are identified. Second, Philadelphia contained large enough numbers of blacks living in the suburbs to reliably estimate (1) separate black and white means by area for our principal measure of travel time, and (2) the effect of travel time on the employment of all four black youth groups considered below.

An obvious measure of the expected commuting time of the marginal youth is the mean travel time of the area's youth who have jobs. Mean travel times of youth (TY) were therefore computed for workers aged 16 to 24 employed in low wage jobs (wage $\leq$ \$5.00/hour). Higher wage youth workers were excluded to improve the measurement of the expected commute of the representative youth. A possible limitation of TY is that it may understate the expected commute in areas with poor job access if working youth tend to be those who were able to find nearby jobs. To mitigate this problem, the mean travel time of all low wage workers (TA) was alternatively employed as a job access measure, based upon the assumption that adults find a long commute less of a problem.[4]

TA was computed separately for whites and blacks and each youth was then assigned one of the 52 values of time depending on his (her) area of residence and race. TY was computed separately for whites and for all workers regardless of race. White youth were assigned the white means, while black youth were assigned the means based on all races. The latter was necessary since black youth travel time means could not be reliably estimated because many of the suburban areas contained too few observations.

A possible problem with the use of average travel time to measure job proximity is that the mix of transportation modes is not uniform among areas. In Daniel McFadden's (1974) comprehensive analysis of urban travel demand, distance to work was found to have a very low correlation with the decision to drive or take a bus to work. This suggests an attempt should be made to standardize for speed of travel by restricting the computation of mean travel times to those workers using a particular mode of transportation. Thus our preferred measure of job access is the average travel time of workers using privately owned motor vehicles. We experimented, however, with travel times inclusive of all modes of travel and the results are robust with respect to alternative measures.

Separate equations for white and black youth were estimated for the following four groups: (1) 16–19-year-olds, living at home, who were enrolled in school; (2) 16–19-year-olds, living at home, who were not in school; (3) 20–24-year-olds, living at home, who were not enrolled and who had less than a college education; (4) 20–24-year-olds, not living at home, who were neither enrolled nor in the military and who had less than a college education.

To control for other factors that might effect the probability of employment, we selected individual (I) and family background (F) variables that have been found to be important in prior studies of youth employment, for example, Richard Freeman (1982), and are available in the PUS. The I variables were age, sex, years of education, health status, whether the high school diploma had been obtained, and marital status. The F variables were family income net of the youth's earnings and the following characteristics of the household head: sex, educational level, occupational class, and employment status. The full set of I and F variables enter the equations estimated for

[3] The PUS also has a 1 percent (*B* Sample) sample. Since the same geographical areas are defined on the *A* and *B* Samples for the Philadelphia SMSA, they were combined to form a 6 percent sample. For Los Angeles and Chicago, the *A* Sample identifies only 3 and 7 areas, respectively, and thus for these two SMSAs the two samples could not be combined.

[4] Setting the maximum wage at \$5.00 in effect limited the sample to workers employed in youth-intensive occupational classes; that is, sales, clerical, service, operatives, and laborers. Ninety-four percent of the employed teenagers in our sample worked in these jobs. The sample of workers used to compute travel times by area included 11,000 observations.

teenagers.[5] A subset of these variables were used for the older, not at home youth: age, years of school, high school diploma, sex, health status, income net of the youth's earnings, and marital status. For consistency these same variables entered the equations estimated for the older at home group.[6]

To measure the relative influence of job accessibility, individual characteristics, and family background on youth employment, we estimated multinomial and dichotomous logit models and ordinary least squares regressions. Multinomial logit was selected to allow for the joint endogeneity of the decisions to attend school and to work. Since the multinomial and dichotomous logit results were virtually identical with regard to the travel time variables, only dichotomous logit and OLS results are present here.[7,8] To conserve space, coefficients on the control variables are not presented, but the results are analogous to those obtained in previous work.

Table 1 presents estimated partial derivatives of job probability with respect to TA and TY obtained from both OLS and logit regressions.[9] Without exception, TA provided a better fit than TY, but both measures yielded similar partial derivatives. In all 32 cases (8 groups times the two job access measures times the two estimating techniques), the estimated travel time partial derivatives are statistically significant and are nontrivial in magnitude. The sizes of the partial derivatives are similar among groups except for the relatively larger effect estimated for older white youth no longer living at home. That the measure of job access would have its largest impact on this group was an unexpected result, since this group is likely to be more mobile and have a greater incentive to acquire information on the location of jobs. Attempts to resolve this paradox lead us to conclude that the estimated partial derivatives for this group were biased upward by self-selection. Once they have obtained a job, older white youth living on their own likely select higher quality residential neighborhoods, which in Philadelphia tend to have lower mean travel times.[10]

The similarity in the sizes of the estimated partial derivatives between enrolled and non-enrolled youth (both old and young) was also somewhat of a surprise, since we anticipated that non-enrolled youth should be more willing and able to commute farther distances. The results suggest that either this is not true or, what is more likely, that poor job access limits the labor market information of all youth, regardless of their enrollment status.

To determine the economic significance of our results, we computed the decline in job probability resulting from a one standard

---

[5] The equation estimated for black in-school teenagers also included a central city dummy variable, since for this group the central city was found to be an outlier with respect to the travel time variables. Including the city dummy, which was positive and significant, increased the efficiency of the estimated time coefficients. The positive city effect may be attributable to less consumer discrimination within the city, since in-school teenagers most frequently work in service jobs. Alternatively, there may have been employment programs targeted to in-school city black youth during the survey period. For all of the other groups, both black and white, the city dummy had little effect on estimated travel time partial derivatives and was insignificant. To reduce multicollinearity, it was excluded from estimated regression equations.

[6] Equations including the full set of variables were also estimated for this group and the estimated partial derivatives of time were virtually the same as those reported in Table 1.

[7] We have also estimated a multinomial logit model for Chicago's teenagers, which included a set of residential location dummies in lieu of a travel time variable. The results are consistent with those reported below in Tables 1 and 2, in that both blacks and whites were found to have significantly higher job probabilities at those locations offering superior job access (Keith Ihlanfeldt and David Sjoquist, 1988).

[8] Both logit and OLS results are presented because we have no a priori expectation regarding which functional form is more nearly correct. Although logit is used with increasing frequency for estimating equations with dichotomous dependent variables, Thomas Stoker (1986) has shown that OLS may be more appropriate in a broad variety of circumstances.

[9] In the logit case, implied partial derivatives are calculated at the mean values of the independent variables.

[10] Due to the limitations of the data, we were unable to control for possible simultaneity between the probability of having a job and travel time.

TABLE 1—ESTIMATED PARTIAL DERIVATIVES OF JOB PROBABILITY WITH RESPECT TO MEAN TRAVEL TIME: PHILADELPHIA[a]

	Time Variables[b]				Employment Rate	Percent of Racial Difference Explained[c]		Sample Size
	TA		TY					
	OLS	Logit	OLS	Logit		OLS	Logit	
16–19-Years Old, Enrolled								
Whites	−0.016	−0.017	−0.012	−0.013	0.35	51	54	10,418
	(7.59)	(7.87)	(6.79)	(7.00)				
Blacks	−0.013	−0.011	−0.013	−0.011	0.12	42	35	2,800
	(3.36)	(3.17)	(1.92)	(1.89)				
16–19-Years Old, not Enrolled								
Whites	−0.017	−0.020	−0.013	−0.013	0.67	33	39	3,519
	(4.90)	(4.92)	(4.63)	(4.69)				
Blacks	−0.018	−0.017	−0.020	−0.019	0.28	35	33	1,182
	(4.81)	(4.61)	(3.64)	(3.57)				
20–24-Years Old, at Home								
Whites	−0.015	−0.018	−0.010	−0.009	0.78	31	38	6,615
	(6.46)	(6.71)	(5.09)	(5.43)				
Blacks	−0.014	−0.016	−0.018	−0.020	0.42	29	33	2,253
	(4.87)	(4.97)	(4.02)	(4.22)				
20–24-Years Old, not at Home								
Whites	−0.020	−0.023	−0.015	−0.019	0.68	71	81	6,253
	(7.91)	(8.10)	(6.68)	(6.77)				
Blacks	−0.010	−0.014	−0.013	−0.019	0.49	36	50	1,405
	(3.32)	(3.67)	(2.69)	(3.11)				

[a] Asymptotic t-statistics in parentheses for logit and t-statistics for OLS.

[b] TA = Mean travel time of low-wage workers using private motorized transportation.
TY = Mean travel time of low-wage workers aged 16–24 using private motorized transportation.

[c] The percentages reported for whites (blacks) were obtained by assigning white (black) youth the black (white) mean TA value of the group and using the estimated partial derivatives from the white (black) equations.

deviation increase in TA (2.2 minutes for whites, 3.7 minutes for blacks) suggested by the partial derivatives from the logit equations as presented in Table 1. For whites the increase in time reduced job probability by 3.8 to 5.1 percentage points, depending on the group considered. The corresponding range for blacks was 4.0 to 6.3 percentage points. An even more dramatic illustration is obtained by predicting the change in employment probability if a youth moved from the worst to the best job access area. For whites the increase is 12.6 to 17.0 percentage points, depending on the group considered, and for blacks it is 14.4 to 22.3 percentage points. These results stand in sharp contrast to the findings of Ellwood and Leonard. Ellwood found that a one standard deviation change in his best measure of job access (described below) causes a change in job probability of 0.5 percentage points, while a one standard deviation change in Leonard's job access measure changed job probability by 1.3 percentage points.

In Philadelphia, black youth have decidedly worse access to employment opportunities than white youth: the mean value of TA is 26.0 minutes for blacks but only 18.5 minutes for whites. A portion of the difference in black and white employment rates can therefore be attributed to differential job access. Estimates of the magnitude of this portion for each enrollment and age group were obtained by conducting a partial decomposition analysis (Ronald Oaxaca, 1973; Alan Blinder, 1973). This involved predicting the job probability for a black youth with black mean values of all characteristics, represented by $\bar{X}$, but with the same accessibility to jobs as the average white youth:

$$\hat{P}_B = \hat{a}_B + \hat{b}_B \bar{X}_B + \hat{c}_B \overline{\mathrm{TA}}_W, \tag{1}$$

where W and B refer to the white and black samples, respectively. The difference between $\hat{P}_B$ and the actual employment rate for white youth (P_W) yields an estimate of the racial difference in employment rates that would exist if blacks and whites had identical job access. This hypothetical difference in employment rates is subtracted from the actual difference in employment rates, $(P_W - P_B) - (P_W - \hat{P}_B) = \hat{P}_B - P_B$, and expressed as a percentage of the actual difference. Equivalent decompositions were also performed for white youth. Excluding the estimates based on the partial derivatives which we believe are biased by self-selection, the range in the amount of the racial gap in employment rates attributable to job access was 33 percent to 54 percent and 29 percent to 51 percent using the logit and OLS results, respectively (see columns 6 and 7 of Table 1). These estimates strongly indicate that racial differences in the nearness of jobs is important in explaining the relatively low employment rate of all black youth, regardless of their age or enrollment status.

As a check on our results, we conducted a specification test that involved substituting a full set of location or area dummy variables for the commuting time variable in the OLS employment equations for blacks. Using these equations, we predicted the black employment rate under the assumption that blacks are distributed across areas in the same proportion as whites. This was accomplished by multiplying the estimated area dummy coefficients by the white distribution across areas. A decomposition was conducted analogous to the one performed with travel time.

If blacks, when distributed across areas in the same proportion as whites, have essentially the same access to jobs as whites, then the decomposition based on the area dummies and on travel time should yield similar results in the absence of serious specification errors.[11] However, different results will be obtained if the travel time variable is correlated with excluded locational factors because the area dummies capture the effect of time as well as other locational factors. The two decompositions yielded virtually identical results. For example, as reported in Table 1, the OLS travel time decomposition for out-of-school teenagers explained 35 percent of the racial difference in employment rates, while the area dummies decomposition explained 34 percent.[12] The two decompositions yielded nearly equal percentages for the other groups.

While the results presented for Philadelphia provide strong support for the job access hypothesis, there remains the possibility that Philadelphia is a unique case. We now turn to evidence for other SMSAs which suggests this is not true.

II. Results for Other Metropolitan Areas

The B Sample of the PUS, but not the A Sample, provides sufficient spatial disaggregation to investigate the relevance of the job access hypothesis as it applies to black youth joblessness in Los Angeles and Chicago.

---

[11] When blacks are allocated to areas in the same proportion as whites, there are two effects on travel time that are captured by the decomposition based on travel time—an inter-area difference in race-specific travel time and an intra-area difference equal to the racial difference in travel time. The decomposition based on area dummy variables captures only the inter-area difference. For the case of blacks being allocated to the white world, the intra-area difference in travel time is very small and, therefore, the two decompositions should yield similar results.

[12] The specification test was also performed by predicting the white employment rate if whites were distributed across areas in the same proportion as blacks. For this case the intra-area difference in travel time is large. The reason for this is that racial differences in travel time are much smaller in the suburban areas than in the central city where the greatest percentage of blacks reside. We, therefore, expect that the decomposition based on area dummies would explain a smaller percentage of the racial gap in employment rates than the decomposition based on travel time (see fn. 11). For example, the former decomposition explained 13 percent of the gap for out-of-school teenagers, while the latter decomposition, as reported in Table 1, explained 33 percent.

TABLE 2—ESTIMATED PARTIAL DERIVATIVES OF JOB PROBABILITY WITH RESPECT TO MEAN TRAVEL TIME: CHICAGO AND LOS ANGELES[a]

	Chicago			Los Angeles		
	TA		Sample Size	TA		Sample Size
	OLS	Logit		OLS	Logit	
Black Youth						
16–19, Enrolled	−0.019[b]	−0.016[b]	862	−0.024[b]	−0.021[b]	556
	(3.31)[a]	(3.30)		(2.28)	(2.23)	
16–24, Non-Enrolled at Home	−0.014[b]	−0.014[b]	835	−0.055[b]	−0.060[b]	484
	(1.84)	(1.76)		(3.74)	(3.51)	
16–24, Non-enrolled not at Home	−0.023[b]	−0.030[b]	453	−0.026[b]	−0.031	368
	(3.03)	(2.90)		(1.65)	(1.60)	
White Youth						
16–19, Enrolled	−0.009[b]	−0.009[b]	2,416	0.005	0.002	2,034
	(2.45)	(2.65)		(0.60)	(0.30)	
16–24, Non-enrolled at Home	−0.005[b]	−0.005	2,131	−0.014[b]	−0.013	1,864
	(1.70)	(1.43)		(1.70)	(1.55)	
16–24, Non-enrolled not at Home	−0.001	−0.002	1,882	−0.002	−0.004	2,319
	(0.11)	(0.70)		(0.11)	(0.45)	

[a] Asymptotic t-statistics in parentheses for logit and t-statistics for OLS.
[b] Significant at the 5 percent level, one-tailed test.

There are 28 and 32 areas identified for Los Angeles and Chicago, respectively. For each of these areas, TA was computed for whites and for all workers, regardless of race.[13] White youth were assigned the mean white travel times and black youth were assigned the mean times computed for all races. (The number of black observations outside the central city was too small to reliably estimate separate black mean travel times.)

Because of the smaller number counts of black youth provided by the *B* Sample, partial derivatives of job probability with respect to TA were computed for only the following three groups: (1) 16–19-year-olds, living at home, and enrolled; (2) 16–24-year-olds, living at home, non-enrolled, and with less than a college education; and (3) 16–24-year-olds, not living at home, non-enrolled, and with less than a college education. The control variables were exactly the same as those employed for Philadelphia, with the full set of individual and family background variables included for youth living at home and a subset of these variables used for youth living on their own. Estimated partial derivatives, again obtained from both logit and OLS regressions, are reported in Table 2. Since the estimates are very similar, we will refer only to the logit results.

Sample sizes are much smaller than those employed for Philadelphia, which perhaps explains the greater range in the magnitudes of the estimated partial derivatives among the various youth groups. Nevertheless, for black youth five of the six partial derivatives are significant at the 5 percent level, and the remaining partial derivative is nearly significant at this level. In all cases the magnitude of the effect of job access on job probability suggested by the size of the partial derivative is again substantial, with a one standard deviation increase in TA increasing employment by 3 to 9 percentage points, depending on the group considered. This range is similar to the 4-to-6 percentage point range obtained for Philadelphia.

For white youth, the partial derivative estimated for Chicago's enrolled 16–19-year-olds was −0.009 and highly significant.[14] In the other five cases the estimated partial derivatives for white youth were not significant at the 5 percent level but two were significant at the 10 percent level. Within each youth group, the partial derivative was much smaller for white youth in comparison to black youth in both SMSAs.

---

[13] Since TY and TA yielded similar results for Philadelphia we focus on TA. In addition, the number counts for TY using the *B* Sample were small.

There are two possible explanations for the contrasting results obtained for white and black youth. First, in contrast to Philadelphia, white mean travel time in Los Angeles displayed considerably less variation. The variance in time was 4.84 minutes in Philadelphia, but only 1.69 minutes in Los Angeles. It is therefore possible that intrametropolitan differences in job access are simply too small for white youth living in Los Angeles to have much of an effect on their job probability. However, the variation in times was as large in Chicago as in Philadelphia. This suggests that in Chicago, and perhaps also in Los Angeles, poor job access is more easily overcome by white youth in comparison to black youth. Whites may have better informal job networks that make them less dependent than blacks on jobs available within the immediate area.

In addition to the results presented in Table 2, additional experimentation was conducted for Chicago in an attempt to reconcile our results with those of Ellwood. His sample included non-enrolled youth of all races, 16 to 21 years of age. The employment rate of these youth within the census tract was regressed on the average time to work of workers living in the youth's community (defined as groups of census tracts) and on variables describing the socioeconomic characteristics of the tract's residents. Mean travel times for all workers, teenagers, and blue-collar plus service workers were alternatively employed. The latter time variable performed best, but nevertheless indicated that job access has a trivial effect on youth employment.

Using the same sample definition as Ellwood and similar control variables (at the individual, not tract level), we obtained a partial derivative of job probability of −0.011 (standard error = 0.003) with respect to TA. When travel times of blue-collar and service workers were substituted for TA, the equation fit the data less well and a partial derivative of −0.008 (standard error = 0.002) was obtained. However, even in the latter case, the partial derivative is substantially larger than the −0.001 estimate reported by Ellwood. These results, along with those presented in Table 2, suggest that Ellwood's much smaller estimated effect may be the result of four factors: (1) his use of the employment rate of youth, regardless of race, as his dependent variables; (2) his use of tract-level rather than individual-level dependent and control variables; (3) his use of an inferior measure of the expected commuting time of the marginal youth; and (4) as expressed by Leonard (1986b), the possibility that Ellwood's employment rates and mean travel times may have been unreliable because they were based on small samples.

III. Conclusion

In his recent essay on youth joblessness, Albert Rees (1986) lists three hypotheses that might explain the high rate of joblessness of young blacks relative to young whites: (1) it arises from certain aspects of the culture of the black ghetto, (2) it is the result of employer discrimination in hiring against young blacks, and (3) it is due to jobs moving away from the inner city, leaving blacks with poor access to employment opportunities in comparison to those available to whites. While not ruling out the possible importance of the first two explanations, this paper is the first to offer evidence supporting the validity of the job access hypothesis. The nearness of jobs was found to have a strong effect on the job probability of both white and black youth living in Philadelphia and differential job access was found to explain a large portion of the racial difference in youth employment rates. Results for other areas confirm the results obtained for Philadelphia black youth, but suggest the possibility that the affect of job access on white youth may vary among metropolitan areas.

---

[14] An equation for Chicago's non-enrolled white teenagers was also estimated. The implied partial derivative of job probability with respect to travel time was also −0.009 for this group and was highly significant.

An important issue left unanswered by the present study is exactly how an absence of nearby jobs impinges on black youth employment. Blacks living in neighborhoods with low job proximity may be unemployed because they have limited information on job opportunities, or they may experience difficulty commuting to places where jobs are available, or they may be reluctant to search for jobs in unfamiliar areas. Knowledge of the relative importance of these various factors is crucial to establishing effective policies to combat the black youth employment problem.

APPENDIX: SAMPLE SELECTION AND VARIABLE CONSTRUCTION

All variables are from the Census of Population and Housing, 1980: Public-Use Microdata Samples and are referred to below by the names, in capitalized letters, used in the PUS *Data Dictionary*. For Philadelphia, we combined the *A* and *B* sample, while for the other areas the *B* Sample was used.

Sample Selection. The various samples were created by first selecting persons of the appropriate ages (AGE), school enrollment (SCHOOL), and whether living at home. At home status was determined by using RELAT1. A youth was classified as living at home if the youth's value for RELAT1 was 02, 03, or 05, and not at home if RELAT1 was 00, 01, 06, 07, or 09. There were no observations with RELAT1 = 04. Observations with RELAT1 = 08, 10, or 11 were excluded. Youth in the armed services (LABOR = 4 or 5) were excluded.

The SMSA and areas within the SMSAs were determined using SMSA and COGRP.

Construction of Variables. The following variables were used directly from the PUS for the person (youth or head as appropriate): AGE, SEX, RACE (01 and 02 only), MARITAL, and DISABIL1. Years of education completed and existence of high school diploma were determined by using GRADE and FINGRADE. Family income was constructed as the family income (FAMINC) less income (INCOME1) for the youth. Both income and income squared were included in the regressions.

To identify the head of the household, we used RELAT1. In general the person identified as our head of the household was the census householder, that is, RELAT1 = 00. There were exceptions. If the census householder was female, we searched for a spouse, and if present, he was denoted as our head. If a youth was the census householder, we searched for the presence of a parent, RELAT1 = 04; if there was a male parent, he was designated our head, otherwise the female was. If there was no parent present, we used the adult relative with the highest wage and salary income, INCOME1, as the head.

Five occupational classes were constructed using OCCUP, where Managerial = 1 if $0 < \text{OCCUP} \leq 199$, Technical = 1 if $200 \leq \text{OCCUP} \leq 399$, Service = 1 if $400 \leq \text{OCCUP} \leq 469$, Craft = 1 if $470 \leq \text{OCCUP} \leq 699$, and Labor = 1 if $700 \leq \text{OCCUP} \leq 889$.

Employment status was determined by LABOR: LABOR = 1 or 2 implied employed and LABOR = 3 or 6 implied nonemployed. If LABOR = 0, 4, or 5, the observation was dropped.

The central city dummy equaled 1 if AREATYPE = 1.

To construct the commuting time variables, we first constructed for each subarea (COGRP) a sample of workers with hourly wage rates less than or equal to $5.00 and a value of TIME greater than 0 and less than 60. The wage rate was calculated as INCOME/(HOURS 79 * WEEKSW79). The mean of TIME for each subarea were calculated for two subgroups, youth 16–24 years of age using just private motorized vehicles (TY) and all workers using just private motorized vehicles (TA). A value for MEANS = 01, 02, 03, or 08 implied commuting by a private motorized vehicle.

REFERENCES

Borus, Michael E., "Willingness to Work Among Youth," *Journal of Human Resources*, Fall 1982, *17*, 581–93.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, *8*, 436–55.

Cogan, John F., "The Decline in Black Teenage Employment: 1950–1970," *American Economic Review*, September 1982, *72*, 621–38.

Ellwood, David, T., "The Spatial Mismatch Hypothesis: Are There Teenage Jobs Missing in the Ghetto?," in Richard B. Freeman and Harry J. Holzer, eds., *The Black Youth Employment Crisis*, Chicago: University of Chicago Press, 1986, ch. 4.

Feldstein, Martin and Ellwood, David T., "Teenage Unemployment: What Is the Problem?," in Richard B. Freeman and David A. Wise, eds., *The Youth Labor Market Problem: Its Nature, Causes, and Consequences*, Chicago: University of Chicago Press, 1982, ch. 2.

Freeman, Richard B., "Economic Determinants of Geographic and Individual Variation in the Labor Market Position of Young Persons," in Richard B. Freeman and David A. Wise, eds., *The Youth Labor Market Problem: Its Nature, Causes, and Consequences*, Chicago: University of

Chicago Press, 1982, ch. 5.

Gordon, Robert J., "Comment on: Time Series Changes in Youth Joblessness," in Richard B. Freeman and David A. Wise, eds., *The Youth Labor Market Problem: Its Nature, Causes, and Consequences*, Chicago: University of Chicago Press, 1982, 189–98.

Holzer, Harry J., "Informal Job Search and Black Youth Unemployment," *American Economic Review*, June 1987, *77*, 446–52.

Ihlanfeldt, Keith and Sjoquist, David L., "The Effect of Residential Location on the Probability of Black and White Teenagers Having a Job," mimeo., 1988.

Kain, John F., "Housing Segregation, Negro Employment, and Metropolitan Decentralization," *Quarterly Journal of Economics*, May 1968, *82*, 175–97.

Kasarda, John D., "Urban Change and Minority Opportunities," in Paul E. Peterson, ed., *The New Urban Reality*, Washington: The Brookings Institution, 1985, 33–68.

Leighton, Linda and Mincer, Jacob, "Labor Turnover and Youth Unemployment," in Richard B. Freeman and David A. Wise, eds., *The Youth Labor Market Problem: Its Nature, Causes, and Consequences*, Chicago: University of Chicago Press, 1982, ch. 8.

Leonard, Jonathan S., (1986a) "Space, Time and Unemployment: Los Angeles 1980," unpublished manuscript, 1986.

______, (1986b) "Comments On: The Spatial Mismatch Hypothesis: Are There Teenage Jobs Missing in the Ghetto?," in Richard B. Freeman and Harry J. Holzer, eds., *The Black Youth Employment Crisis*, Chicago: University of Chicago Press, 1986, 185–90.

McFadden, Daniel, "The Measurement of Urban Travel Demand," *Journal of Public Economics*, November 1974, *3*, 303–28.

Oaxaca, Ronald, "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, October 1973, *14*, 693–709.

Rees, Albert, "An Essay on Youth Joblessness," *Journal of Economic Literature*, June 1986, *24*, 613–28.

Stoker, Thomas M., "Consistent Estimation of Scaled Coefficients," *Econometrica*, November 1986, *54*, 1461–81.

Wilson, William Julius, *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*, Chicago: University of Chicago Press, 1987.

U.S. Bureau of the Census, Census of Population and Housing: Public-Use Microdata Sample (*A* and *B* Sample), (machine-readable data file), Washington: Bureau of the Census, 1983.

______, Public-Use Microdata Samples Technical Documentation, prepared by the Data User Services Division, Washington: Bureau of the Census, 1983.

U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings*, *35*, December 1988.

Sex Discrimination in Labor Markets: The Role of Statistical Evidence: Comment

By DEBRA A. BARBEZAT AND JAMES W. HUGHES*

In a recent article in this *Review* (Kuhn, 1987), Peter Kuhn seeks to assess the relative importance of statistical and other types of evidence in determining women's perceived level of discrimination. The author's point of departure is what he characterizes as a "surprising" relationship. He finds a robust, but statistically insignificant, negative correlation between the conventional measure of salary discrimination ("statistical evidence," in Kuhn's terminology) facing a female worker and the probability that the worker will report discrimination on a confidential survey. In other words, workers facing higher levels of measured salary discrimination are less likely to perceive discrimination.

Kuhn explains this finding by hypothesizing the existence of some other evidence ("nonstatistical evidence"), which, in addition to the salary differential, leads women to recognize and report sex discrimination in their employment. If the quantity of nonstatistical evidence is low, according to Kuhn, women may not report discrimination even if the statistically measured salary differential is large. Kuhn assumes that the quantity of nonstatistical evidence, which is observed by the worker, but not the researcher, is a function of the worker's observed human capital characteristics. Finally, Kuhn evaluates the effect of the nonstatistical evidence using the worker's observed characteristics as a proxy for the unobserved nonstatistical evidence.

According to Kuhn's findings, the unobserved variable is a more important determinant of the worker's decision to report discrimination than is the level of statistically measured wage discrimination. Kuhn uses these results to question the importance of recent Supreme Court decisions allowing statistical measures of discrimination to be admitted into evidence in civil sex discrimination suits. As he concludes that such evidence is not the primary factor influencing workers' reporting behavior, allowing these measures to be presented in court should not lead to a large increase in litigation.

In this comment, we argue that Kuhn's result, that workers facing a high level of measured discrimination are less likely to report the discrimination, may be explained by incorporating the quality of information, employers' preferences, and the costs of discrimination into the model. We derive conditions under which the negative correlation between measured and reported salary discrimination may be seen as the result of a rational balancing by employers of the benefits and costs of discrimination. Briefly, employers are more likely to discriminate when employees have less accurate information and the probability of reporting, and hence detection, is low. We also offer statistical evidence from the academic labor market to illustrate this alternative hypothesis.

I. The Negative Correlation Between Measured Discrimination and Reported Discrimination

Kuhn's result may be surprising in that one might expect a larger number of discrimination reports from employees of firms that discriminate a great deal. In this section we discuss how differences in the quality of statistical information available to workers influences their decision to report discrimination, and how employers may alter their choice of discrimination in response. This model offers an explanation for the negative correlation uncovered by Kuhn that does not require the existence of nonstatistical evidence of discrimination.

---

*Department of Economics, Amherst College, Amherst, MA 01002. The authors wish to thank Geoffrey Woglom, Peter Kuhn, and an anonymous referee for many helpful comments.

A. *The Quality of Statistical Information*

We assume a female employee is concerned only with the true level of salary discrimination, denoted D_i. The woman's best estimate of discrimination, $\hat{D}_i$, is related to the true level D_i through the following expression,

$$\hat{D}_i = D_i + \epsilon_i. \tag{1}$$

The statistical noise in the worker's estimate of discrimination varies across firms and individuals such that ϵ_i is distributed with mean zero and variance σ_i^2. Following Kuhn, we assume that a woman reports discrimination if she is k percent certain that the true level of salary discrimination, D_i, is greater than zero. The statistical test employed is to report discrimination if

$$\hat{D}_i > z(k_i), \tag{2}$$

where z is the critical value of a one-tailed test of significance level k, testing the null hypothesis that the true level of salary discrimination, D_i, equals zero. The critical value z will depend on the "quality" of the employee's information, σ_i^2. For a given level of salary discrimination D_i, and a test statistic significance level k, employees with more accurate information (i.e., lower σ_i^2) will be more likely to report discrimination compared to employees with less accurate information (i.e., higher σ_i^2). In addition, the probability of reporting discrimination rises with the true value of D_i, but the probability is always lower for employees with poorer quality information.

We illustrate this relationship using a simplified example. Assume ϵ_i has the following uniform distribution

$$f(\epsilon_i) = \frac{1}{2c} \quad \text{for } \epsilon_i \in [-c, +c] \tag{3}$$

$$= 0 \text{ otherwise.}$$

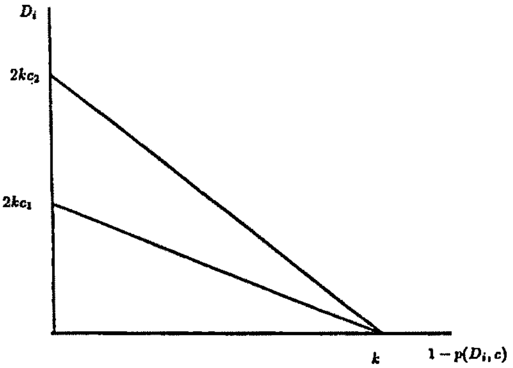


FIGURE 1. PROBABILITY OF REPORTING DISCRIMINATION, D_i AND QUALITY OF STATISTICAL INFORMATION

For a given level of significance k, the critical value z is[1]

$$z = 2ck - c. \tag{4}$$

As the dispersion in the distribution rises (falls) with c, the critical value for reporting discrimination also rises (falls) for levels of significance k greater than 0.5. The employer chooses a level of discrimination D_i. The probability that the employee *does not* report discrimination, $1-p(D_i, c)$, is given by

$$\Pr[\hat{D}_i < z \mid D_i] = \int_{D_i - c}^{z} \frac{1}{2\epsilon} d\epsilon \tag{5}$$

$$= k - \frac{D_i}{2c} = 1 - p(D_i, c).$$

The probability of reporting rises with the true level of discrimination, D_i, and falls as c rises (i.e., the employees' information is of lower quality). Equation (5) is shown graphically in Figure 1, where c_2 represents lower quality information relative to c_1.

B. *Employer Preferences*

If workers vary in their ability to detect salary discrimination, employers may try to exploit this difference to their advantage. In choosing a level of discrimination, employers must weigh the benefits of discrimination (lower production costs) against the expected cost should the practice become public knowledge (lower productivity, civil penalties).[2] The expected cost of the civil penalties component may be thought of as the probability of detection multiplied by the average cost per violation. If, as is likely, the probability of detection is positively correlated with the number of reports filed by employees, the probability of detection, and therefore the expected cost of discrimination, will be higher for firms with better-informed employees.

---

[1] Derivation of the example is shown in Appendix 1.

We may write the employer's utility function as

$$U(D_i,(1-p_i));\ U_1,U_2>0;\qquad U_{11},U_{22}>0, \tag{6}$$

where D_i and p_i are defined in (1) and (5). If we assume the constant elasticity of substitution form for the utility function, (6) becomes

$$U(D_i,(1-p_i)) = \alpha D_i^{\rho}+(1-\alpha)(1-p_i)^{\rho}. \tag{7}$$

The employer's problem is to maximize (7) subject to (5). Substituting equation (5) for $(1-p_i)$ in equation (7), the problem may be written as

$$\max_{D_i} \alpha D_i^{\rho}-(1-\alpha)\left(k-\frac{D_i}{2c}\right)^{\rho}. \tag{8}$$

The first-order condition for this problem, given values of k and c, is

$$\frac{dU}{dD_i}=\rho\alpha D_i^{\rho-1} - \rho(1-\alpha)(2c)^{-1}\left(k-\frac{D_i}{2c}\right)^{\rho-1} = 0. \tag{9}$$

Rearranging, the employer's utility-maximizing choice of D_i, denoted D_i^*, is

$$D_i^* = k\left[(2c)^{-1}+(2cs)^{r}\right]^{-1}, \tag{10}$$

where $r\equiv 1/(\rho-1)$ and $s\equiv\alpha/(1-\alpha)$.

To examine how the employer's optimal choice varies with changes in the quality of the employee's statistical information, we differentiate D_i^* with respect to c, and rearrange as follows:

$$\frac{dD_i^*}{dc}=\frac{D_i^*}{c}\left[(2c)^{-1}-r(2cs)^{r}\right] \times\left[(2c)^{-1}+(2cs)^{r}\right]^{-1}. \tag{11}$$

Equation (11) is positive for all values of $r\leq 0$, which corresponds to all admissible values of $\rho\in[-\infty,1]$. Thus, as the employee's quality of information falls (i.e., c rises), the employer will always choose to increase the level of salary discrimination.

We may consider changes in the probability that the employee will report discrimination by differentiating equation (5) with respect to c at D_i^*.

$$\frac{d(1-p(D_i^*,c))}{dc}=\frac{D_i^*}{2c}-\frac{1}{2c}\frac{dD_i^*}{dc} = \frac{1}{2c^2}\left[D_i^*-c\frac{dD_i^*}{dc}\right]. \tag{12}$$

Equation (12) is positive if $D_i^*>c[dD_i^*/dc]$. Substituting for dD_i^*/dc from (11), we find that equation (12) is positive when

$$\left[(2c)^{-1}-r(2cs)^{r}\right]\left[(2c)^{-1}+(2cs)^{r}\right]^{-1} <1. \tag{13}$$

---

[2] While discrimination may raise labor costs, the firm may also "save" other resources used to monitor and enforce nondiscriminating behavior. See Brown (1982), pp. 40–41 and fn. 11.

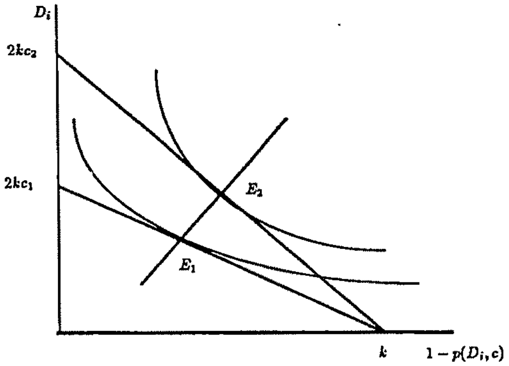


FIGURE 2. EMPLOYER'S UTILITY-MAXIMIZING CHOICE OF DISCRIMINATION D_i AS A FUNCTION OF THE QUALITY OF STATISTICAL INFORMATION

Thus, $d(1-p(D_i^*,c))/dc>0$ whenever $r>-1$, or equivalently, whenever $\rho<0$.[3] In other words, if the employer's elasticity of substitution between salary discrimination and employee discrimination reports is less than one (i.e., $\rho<0$), the level of salary discrimination will *increase*, and the probability of reporting discrimination will *decrease* as the quality of employee statistical information falls. Such a change is shown in Figure 2 by the line passing through E_1 and E_2. Under these conditions, we would expect to observe a negative correlation between statistically measured salary discrimination and reported salary discrimination.

Intuitively, this model implies that employers have a downward sloping demand for discrimination. If employer preferences are such that the chosen level of discrimination in negatively related to its "price" (i.e., the expected cost of detection), we should expect to see *less* measured salary discrimination in firms where the probability of detection is high, and *vice versa*.

While this explanation of the negative correlation between measured and reported discrimination involves an unobserved variable, as does Kuhn's analysis, two differences should be emphasized. First, the analysis presented here explains the negative correlation as arising from the interaction between employee and employer choices. Kuhn's model takes employer choices as given, modeling only employees' reactions to these decisions. Second, the above analysis, while not disproving Kuhn's hypothesis concerning employees' use of nonstatistical evidence, does not require an additional category of information to explain the observed negative correlation between the probability of reporting discrimination and the level of statistically measured salary discrimination. Our model relates only to the quality of the *statistical* evidence itself, rather than to the presence or absence of nonstatistical evidence. These differences motivate our alternative explanation of the negative correlation, which we describe in the following section.

II. Nonstatistical Evidence

Kuhn interprets the negative correlation between measured and reported discrimination as an indication that female workers must employ some other information, in addition to observable salary differentials, when deciding whether to report discrimination. He hypothesizes that either the availability of such information, or the ability to interpret this nonstatistical evidence is a function of the woman's characteristics. Pursuant to this hypothesis, he uses women's characteristics as a proxy for the unobserved level of nonstatistical evidence. He finds, by his interpretation, that the level of nonstatistical evidence is the primary determinant of women's reporting behavior. The effect of nonstatistical evidence appears to be particularly strong for young, well-educated women.

Kuhn's model includes no variable controlling for occupation or any other *employer* characteristic.[4] Occupational type is a primary determinant of workers' wages. In particular, occupational type is highly correlated with female worker's characteristics, especially age and education, which have the strongest explanatory power in Kuhn's results. We suggest that the effects which Kuhn attributes to differences in nonstatistical evidence may simply be the result of differences in the occupations of survey respondents.

---

[3] The equilibrium locus in Figure 2 will be negatively sloped when $0<\rho<1$, vertical when $\rho=0$ (Cobb-Douglas), and positively sloped when $-\infty<\rho<0$.

[4] Kuhn acknowledges that such control might be desirable, but states that it was not possible given the data available to him (Kuhn, fn. 14).

Given the considerable degree of occupational segregation, it is unlikely that wage determination is either identical across occupations, or varies only randomly. We argue that it is also unlikely that the quality of *statistical* evidence available to female workers varies randomly across occupations,[5] but this assumption is also incorporated into Kuhn's model. Differences in the availability of statistical information will generate systematic differences across occupations in workers' ability to detect discrimination, regardless of the presence or absence of any nonstatistical evidence. As shown in the previous section, differences in the quality of statistical evidence, combined with employers' preferences concerning discrimination, may yield the negative correlation between measured and reported discrimination.

III. Evidence from the Academic Labor Market

If systematic differences in the quality of statistical information across occupations is the source of the negative relationship discovered by Kuhn, then controlling for occupation or firm type should reduce or eliminate the size and significance of the inverse correlation. We examine this hypothesis using empirical evidence from a recently released data base, the 1984 national faculty survey conducted by The Carnegie Foundation for the Advancement of Teaching.[6] The random sample of 5057 individuals was drawn from approximately 300 colleges and universities. We restricted our study to full-time employees in the top three academic ranks (assistant, associate, and full professors) at four-year institutions.[7]

The survey addressed many diverse issues of concern to American academics, including affirmative action. We selected one such question to serve as an indication of the individual's perception of equitable treatment, whether or not the faculty member agreed with the statement: "I am satisfied with the results of affirmative action at this institution" (question 52).[8] We constructed a dummy variable from this question, equal to zero if the respondent agreed with the statement, and equal to one if the respondent disagreed. After deriving estimates of the degree of salary discrimination by sex, we compare responses to question 52 with salary discrimination measures.

We estimate separate salary regressions for male and female faculty members using the logarithm of salary as the dependent variable. The independent variables include: post-degree experience, post-degree experience squared, age, and dichotomous variables representing sex, race, type of contract (nine month or eleven month), receipt of an advanced degree, whether the respondent was primarily engaged in teaching or research, and whether the faculty member spent a relatively large amount of time in administrative work. We also include dummy variables representing article and book publication. We present these results in Table 1,

---

[5] For example, in some occupations, such as civil service, workers are compensated according to a well-defined salary schedule. In the private sector, similar publicly available salary schedules are less common.

[6] A detailed description of the survey appears in Appendix 2.

[7] The data set includes faculty members from the following Carnegie classifications: Research Universities I and II, Doctoral Granting Universities I and II, and Comprehensive Colleges and Universities I and II. Since we divide the sample according to type of control (private versus public), we also exclude liberal arts colleges, which are exclusively private institutions.

[8] We should note that our measure of reported discrimination is not directly comparable to Kuhn's measure. Question 52 does not ask whether the respondent feels she personally has been a victim of discrimination. Moreover, by mentioning affirmative action, our measure might focus attention on particular forms of discrimination, such as discrimination in hiring or promotion.

TABLE 1—DETERMINANTS OF LOG SALARY IN REGRESSIONS WITH AND WITHOUT ACADEMIC RANK, 1984

	Male	Male + Rank	Female	Female + Rank
Constant	9.89	10.27	9.75	10.17
	(0.037)	(0.046)	(0.060)	(0.079)
Nine Month	−0.107	−0.122	−0.130	−0.134
	(0.012)	(0.011)	(0.025)	(0.023)
Teach	−0.022	−0.024	−0.004	−0.017
	(0.012)	(0.011)	(0.024)	(0.022)
Advanced Degree	0.108	0.063	0.085	0.037
	(0.016)	(0.016)	(0.023)	(0.022)
Experience	0.023	0.011	0.010	−0.34E−5
	(0.002)	(0.002)	(0.004)	(0.004)
Experience Squared	−0.0004	−0.0001	−0.0002	0.83E−4
	(0.00005)	(0.00005)	(0.0001)	(0.0001)
Age	−0.0026	0.0004	0.007	0.0045
	(0.0009)	(0.0009)	(0.0015)	(0.0013)
Book 1	0.028	0.014	−0.028	−0.042
	(0.011)	(0.010)	(0.023)	(0.021)
Book 2	0.045	0.013	0.051	0.010
	(0.017)	(0.016)	(0.041)	(0.038)
Book 3	0.060	0.028	−0.041	−0.055
	(0.021)	(0.020)	(0.067)	(0.067)
Article 1	0.033	0.037	0.046	0.053
	(0.019)	(0.018)	(0.029)	(0.026)
Article 2	0.076	0.077	0.016	0.010
	(0.019)	(0.018)	(0.032)	(0.030)
Article 3	0.057	0.045	0.068	0.066
	(0.019)	(0.018)	(0.030)	(0.028)
Article 4	0.086	0.070	0.145	0.142
	(0.021)	(0.019)	(0.042)	(0.039)
Article 5	0.168	0.128	0.239	0.127
	(0.021)	(0.019)	(0.052)	(0.050)
Black	0.038	0.052	0.195	0.148
	(0.050)	(0.046)	(0.068)	(0.063)
Current Admin.	0.057	0.033	0.099	0.062
	(0.016)	(0.015)	(0.031)	(0.029)
Assistant Professor		−0.219		−0.252
		(0.018)		(0.013)
Associate Professor		−0.142		−0.163
		(0.012)		(0.030)
N	1209	1209	320	320
$\bar{R}^2$	0.48	0.55	0.41	0.50

along with a second regression controlling for academic rank in addition to the above variables.[9]

We use the coefficients from these regressions to calculate the following measure of salary discrimination following Oaxaca (1973) and Kuhn

$$\tilde{D}_i = \hat{\beta}^M X_i^F - w_i^F, \tag{14}$$

where $\hat{\beta}^M$ are the estimated coefficients from the male salary regression X_i^F are the women's characteristics, and w_i^F is the woman's actual salary. $\tilde{D}_i$ measures the difference between the woman's actual salary and the estimated salary she would receive

---

[9] Women's slower rate of promotion may reflect discrimination. It may not be desirable, therefore, to estimate discrimination while controlling for rank.

TABLE 2—PROBIT ESTIMATES OF EFFECT OF MEASURED DISCRIMINATION ON SATISFACTION WITH RESULTS OF AFFIRMATIVE ACTION

	All Institutions		Public Institutions		Private Institutions	
	(1) Basic	(2) Basic + Rank	(3) Basic	(4) Basic + Rank	(5) Basic	(6) Basic + Rank
Constant	−0.029	−0.036	−0.306	−0.042	0.027	0.0002
	(0.076)	(0.074)	(0.083)	(0.081)	(0.197)	(0.189)
D_1	−0.791[a]	−0.890[a]	−0.680	−0.811	−1.20	−1.13
	(0.403)	(0.441)	(0.462)	(0.525)	(0.881)	(0.875)
Log-Likelihood	−219.11	−219.00	−175.22	−175.11	−43.75	−43.84
N	320	320	255	255	65	65

[a] Significant at the 5 percent level.

were her characteristics compensated according to the men's reward structure.[10]

The probit results for the full sample of 320 women are shown in the first two columns of Table 2. In column (1), $\tilde{D}_i$ is measured without controlling for academic rank, while in column (2), $\tilde{D}_i$ is calculated taking academic rank into account. In both cases, there is a significant (at the 5 percent level), negative correlation between measured discrimination calculated using equation (14) and women's satisfaction with affirmative action at their institutions.[11]

Can Kuhn's explanation of the inverse relationship found in his own data explain the similar result found here? If the quantity of nonstatistical evidence is a function of workers' characteristics, we can see no reason why, on average, faculty members' human capital characteristics should differ substantially across these institutions. We hypothesize that differences in the ability to discriminate and differences in detection are more likely to explain the inverse relationship uncovered in the academic labor market.[12]

---

[10] Our discrimination measure, $\tilde{D}_i$, is equivalent to Kuhn's measure D_{1i}. Kuhn also estimates probits for another measure of discrimination, $D_{2i} = X_i^F\hat{\beta}^M - X_i^F\hat{\beta}^F$. Using our data, this measure yielded no statistically significant estimates. Also, please note that while Kuhn assumed that employees observe $\tilde{D}_i$ (Kuhn, equation (7)), we do not assume that employees at either type of institution observe $\tilde{D}_i$.

[11] A similar relationship exists between the survey response and another measure of salary discrimination not examined by Kuhn. We estimate separate regressions for employees of public and private institutions. These regressions include a sex dummy variable (equal to one for males). The coefficient on this dummy variable constitutes a rough estimate of sex discrimination in salaries. We find that this coefficient is consistently larger at the private institutions. The coefficient is equal to 0.065 (indicating approximately a 6.7 percent salary advantage for males) at the public institutions ($N = 1279$, significant at 1 percent). The comparable coefficient for private institutions is twice as large, 0.135 ($N = 300$, significant at 1 percent). When we compare the estimates of salary discrimination to the responses for question 52 (again, our measure of reported discrimination equals zero if the respondent expressed satisfaction with the results of affirmative action at his or her institution, and equals one if the individual disagreed with the statement), we find a negative correlation between measured discrimination and reported dissatisfaction with affirmative action. At research and doctoral institutions the means of this variable were always larger for the public employees, indicating that they were more likely to disagree with the statement. This was the case whether we considered men and women together or the female respondents alone. However, in the third category, comprehensive colleges and universities, private employees tended to have a higher mean than their public counterparts, suggesting a positive relationship between measured discrimination and dissatisfaction with affirmative action at these schools.

[12] If women at private institutions have less accurate information on which to base their reports of discrimination, the error term in the combined public/private

A partial test of this hypothesis may be performed by splitting the sample according to whether the institution is publicly or privately operated. At public institutions, faculty members are state employees. Salary information, which is in the public domain, is often widely circulated. Therefore, the quality of statistical evidence of salary discrimination is likely to be quite good, on average, at these institutions. At private institutions, departments are often excluded from the salary determination process, and salary information is not publicly available. Employees at private institutions are likely to have very poor information, on average, from which to estimate salary discrimination.

If the negative coefficient in column (1) of Table 2 is the result of differences in informational quality at public and private institutions, the size and significance of the coefficient should decline when the sample is split according to type of institutional control (public versus private). When we split the sample and estimate the equations again, we obtain the results shown in columns (3) through (6) of Table 2. In both specifications of the model, the coefficients for $\tilde{D}_i$ remain negative for both private and public institutions, although the estimates were not significant at the 10 percent level for either type of institution. The coefficients for public institutions become smaller in absolute terms compared with the full sample, consistent with our hypothesis. However, the coefficients on $\tilde{D}_i$ for private institutions *rise* substantially, indicating a more pronounced negative relationship, contrary to our hypothesis.[13] These results may indicate that dividing the sample according to institutional control did not fully control for differences in informational quality across employees. Another alternative is that employees observe some nonstatistical information which is more important than statistically measured salary discrimination in the employee's decision to report discrimination. The latter conclusion is consistent with Kuhn's hypothesis.

While our evidence cannot refute Kuhn's hypothesis, our alternative explanation of the negative correlation found in our data and in Kuhn's data has very different implications. Kuhn uses his empirical estimates to conduct simulations of women's reporting behavior when measured discrimination is reduced to zero. Kuhn finds that elimination of measured discrimination would not greatly reduce the number of discrimination reports. However, in our model, where the level of measured discrimination is not exogenously given, but is a choice variable for employers, reducing measured discrimination to zero is only a partial equilibrium experiment. Setting measured discrimination to zero may induce changes in other explanatory variables that may, in turn, significantly, alter the probability of reporting discrimination.

IV. Summary

We believe that Kuhn's finding of a negative correlation between conventionally measured wage discrimination and the probability that women will report discrimination is subject to at least one other interpretation. Employers are more likely to indulge in discrimination when there is a low probability

---

probit regression may be heteroscedastic, as the error variance may be larger for private institutions. The possibility of heteroscedasticity reinforces our point that the negative correlation between measured and reported salary discrimination is the result of incorrectly pooling workers with different quality information. When the sample is split into public and private subsamples, the error terms are homoscedastic.

[13] The result that the negative coefficient becomes larger for private institutions and smaller for public institutions may have an interesting interpretation of its own. Assume that splitting the sample does not perfectly control for the quality of statistical information, but the average level of information is better at public institutions than private institutions. The larger coefficient for private institutions suggests that as the quality of statistical information falls, the chosen level of discrimination rises by a larger proportion. The "demand" for discrimination is perhaps elastic with respect to the probability of detection for low values of the latter. This result suggests that the "offer curve" shown in Figure 2 may not be linear, but, may in fact increase at an increasing rate from left to right.

that such action will be detected. Thus, a high level of measured discrimination will be accompanied, in equilibrium, by a relatively small number of reports, and *vice versa*. If workers have access to *statistical* evidence of varying quality across occupations, or even within a single company, we may observe this negative correlation.

Similar to Kuhn, we have found that a high level of measured discrimination tends to be associated with a relatively low level of discrimination reports. Kuhn might argue that the women's level of nonstatistical discrimination was low, leading to the observed low level of reports. By contrast, our hypothesis indicates that a low level of reported discrimination and a high level of measured discrimination might be cause for greater concern. Women lacking accurate statistical information are not sufficiently confident to conclude that they are being discriminated against. These women will tend not to report discrimination. Moreover, employers may be choosing high equilibrium levels of salary discrimination in situations where employees have poor information concerning colleagues' salaries.

APPENDIX 1:

Derivation of the Probability of Reporting Discrimination

The disturbance in the employee's estimate of discrimination has the distribution

$$\text{(A1)} \qquad f(\epsilon_i) = \frac{1}{2c} \quad \text{for all } \epsilon_i \in [-c, +c]$$

$$= 0 \text{ otherwise.}$$

The probability that the employee does not report discrimination under the null hypothesis D_i equals 0 is

$$\text{(A2)} \quad \Pr\{\hat{D}_i < z | D_i = 0\} = \int_c^z \frac{1}{2\epsilon} d\epsilon = \frac{1}{2} + \frac{z}{2c}.$$

The critical value z is found explicitly by setting (A2) equal to the chosen level of significance k and solving for z

$$\text{(A3)} \qquad \frac{z}{2c} = k - \frac{1}{2} \quad \text{and } z = 2ck - c.$$

For any level of discrimination D_i chosen by the employer, the probability of an employee not reporting discrimination is

$$\text{(A4)} \qquad \Pr\{\hat{D}_i < z | D_i\} = \int_{D-c}^{z} \frac{1}{2\epsilon} d\epsilon$$

$$= \frac{z - D_i + c}{2c}.$$

However, $z/2c = k - 1/2$ from (A3), so (A4) may be written

$$\text{(A5)} \qquad 1 - p(D_i, c) = k - \frac{D_i}{2c}.$$

APPENDIX 2:

Data Set and Variable Construction

The data are taken from the faculty survey portion of the 1984 *Carnegie Foundation National Surveys of Higher Education*. Data requests must be sent to The Carnegie Foundation for the Advancement of Teaching (Princeton, New Jersey). Opinion Research Corporation has prepared a Technical Report describing the data collection process and presenting descriptive statistics for the data base. This report is available through The Carnegie Foundation

We restricted our analysis to full-time faculty members in the top three academic ranks. Full-time status was derived from question 61 (Q61Q equal to 1) and academic rank was based on question 1 (Q1Q equal to 2, 3, or 4). We also limited our sample by selecting faculty who were employed by Research Universities, Doctoral-Granting Universities, and the Comprehensive College and University category (responses 1–6 of question 75, CARN75).

The sources of the variables used in the analysis are described below. Each variable is listed as described in the text, followed by the corresponding question number from the survey and the variable name from the data tape. All variables were created using SPSS-X. All statistical analysis was performed using LIMDEP, by William Greene of New York University.

Logarithm of Salary: The natural logarithm of the values in question 62, Q62Q1-Q62Q2. Interval midpoints were used to derive the values.

Post-degree Experience: Equal to years since receipt of highest degree (Year received highest degree given by question 70, Q70Q).

Age: Question 69, Q69Q. Equal to 84 minus the value given in Q69Q.

Sex: Question 73, Q73Q.

Race: Question 72, Q72Q.

Type of Contract: Question 63, Q63Q.

Receipt of Advanced Degree: Question 21, Q21Q12-Q21-Q22.

Teaching Versus Research Interest: Question 9, Q9Q.

Administrative Work: Question 6, Q6Q4.

Article and Book Publication: Questions 10 and 11, Q10Q and Q11Q.

Academic Rank: Question 1, Q1Q.

Public versus Private Institutional Control: Question 76, CONTROL.

Dummy Variable for Reported Discrimination: Equal to 0 when respondent agreed with question 52 (i.e., Q52Q3 was equal to 1 or 2), otherwise equal to 1.

REFERENCES

Barbezat, Debra A. "Affirmative Action in Higher Education: Have Two Decades Altered Salary Differentials by Sex and Race?," *Research in Labor Economics*, 1989, *10*, 107–56.

Brown, Charles C., "The Federal Attack on Labor Market Discrimination: The Mouse That Roared?," *Research in Labor Economics*, 1982, *5*, 33–68.

Kuhn, Peter J., "Sex Discrimination in Labor Markets: The Role of Statistical Evidence," *American Economic Review*, September 1987, *77*, 567–83.

Oaxaca, Ronald L., "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, October 1973, *14*, 693–709.

1984 Carnegie Foundation National Surveys of Higher Education, Princeton, NJ, The Carnegie Foundation for the Advancement of Teaching, 1984.

Sex Discrimination in Labor Markets: The Role of Statistical Evidence: Comment

By WILLIAM E. EVEN*

In a recent article in this *Review*, Kuhn (1987) provides a description of the link between self-assessed and regression ("statistical") measures of discrimination. This is an important study since courts have formally approved regression estimates as evidence in wage discrimination cases. Undoubtedly, the weight the courts place on such evidence will depend partially upon whether regression estimates are a good proxy for the population's general perception of wage discrimination. The evidence Kuhn presents suggests that regression estimates are not a very good proxy. His major conclusion is as follows (p. 567): "Statistical evidence is found to be less important than other, 'nonstatistical' evidence. The estimates indicate that the percentage of women who report discrimination would fall from 15.4 to 10.2 percent in Canada, and from 9.9 to 6.3 percent in the United States if all conventionally measured pay discrimination were absent." The objective of this comment is to show that Kuhn's conclusions are the result of an arbitrary normalization of an unidentified parameter in his econometric model. If the normalization is adjusted appropriately, *any* conclusion may be made. To illustrate this point, some of the essential features of his model need to be reviewed.

In order to define the statistical measure of pay discrimination, the following wage equations are needed:

$$(1) \qquad w_{ij}^F = \sum_j X_{ij}^F \beta_j^F + \mu_i^F,$$

$$(2) \qquad w_{ij}^M = \sum_j X_{ij}^M \beta_j^M + \mu_i^M,$$

where the superscripts stand for female and male, w_i^F and w_i^M are the logs of individual i's wages, the X_{ij}s include human capital variables, and X_{i0} is a constant term throughout the analysis.

The "statistical" measure of discrimination for individual i is given by:[1]

$$(3) \qquad D_i = \sum_j X_{ij} \hat{b}_j, \qquad \text{where } \hat{b}_j \equiv \hat{\beta}_j^M - \hat{\beta}_j^F.$$

In Kuhn's "naive" model, a probit model is estimated with the dichotomous dependent variable indicating whether a woman reports sex discrimination (an affirmative response is coded as $\mathrm{RD}_i = 1$). The explanatory variables include a constant term and D_i. Kuhn finds the puzzling result that increases in statistical discrimination decrease the probability that a woman will report discrimination. This puzzle, he correctly notes, could be the result of D_i and other "nonstatistical" evidence being a function of the same X_{ij}s. This nonstatistical evidence would include (p. 574) "differences in nonwage benefits, the type of treatment received on the job, as well as other information about wage discrimination not embodied in D_i." The determinants of nonstatistical evidence are defined as:

$$(4) \qquad \omega_i = \sum_j \theta_j X_{ij}^F + \varnothing_i^1,$$

where $\varnothing_i^1$ is a normally distributed error term.

Kuhn defines total discrimination as $(\alpha D_i + \omega_i)$ where α (positive or negative) repre-

---

*Department of Economics, Miami University, Oxford, OH 45056. I would like to thank Peter Kuhn for his assistance in replicating the original results, and Bryce Kanago for comments that helped clarify the discussion.

[1] Kuhn actually has two alternative measures which he labels as D_i^1 and D_i^2. The measure presented in equation (3) is D_i^2 since it is the version he uses in his simulations. (See fn. 18, p. 575 of Kuhn's article.)

TABLE 1—SIMULATION RESULTS

	Case 1[a]				Case 2			
	Canada		United States		Canada		United States	
	($P_1 = 0.155$)		($P_1 = 0.138$)		($P_1 = 0.015$)		($P_1 = 0.007$)	
α/σ_v	P_0	$P_1 - P_0$	P_0	$P_1 - P_0$	P_0	$P_1 - P_0$	P_0	$P_1 - P_0$
−10	0.932	0.777	0.999	0.861	0.864	0.850	0.999	0.992
−5	0.594	0.439	0.867	0.729	0.294	0.280	0.641	0.634
−2	0.30	0.149	0.418	0.279	0.064	0.049	0.091	0.084
−1	0.222	0.0673	0.258	0.120	0.032	0.017	0.029	0.022
0	0.155	0	0.138	0	0.015	0	0.007	0
1	0.103	−0.0522	0.063	−0.075	0.006	−0.009	0.001	−0.006
2	0.0647	−0.0903	0.025	−0.114	0.002	−0.012	0.0002	−0.007
5	0.0116	−0.143	0.00050	−0.138	0.00007	−0.014	5.9E−08	−0.007
10	0.00021	−0.155	2.0E−8	−0.138	2.3E−08	−0.015	2.3E−16	−0.007

Case 3

	Canada		United States	
	($P_1 = 0.640$)		($P_1 = 0.345$)	
α/σ_v	P_0	$P_1 - P_0$	P_0	$P_1 - P_0$
−10	0.914	0.274	0.929	0.584
−5	0.805	0.166	0.704	0.359
−2	0.712	0.072	0.490	0.145
−1	0.677	0.037	0.416	0.071
0	0.640	0	0.345	0
1	0.601	−0.038	0.279	−0.066
2	0.562	−0.078	0.220	−0.125
5	0.442	−0.198	0.091	−0.254
10	0.258	−0.382	0.012	−0.333

Notes: P_1 = Predicted probability of reporting discrimination. P_0 = Predicted probability of reporting discrimination if statistical discrimination is eliminated. $P_1 - P_0$ = Predicted change in probability of reporting discrimination that would result if statistical discrimination is eliminated.

[a] *Case descriptions.* All 3 cases set tenure = 2−3; previous experience dummy = 1 (Canada only); union member, trade school, and physical disability dummies = 0: *Case 1.* Age = 35–44, education = 12, U.S. experience = 10. *Case 2.* Age = 55–76, education < 12, U.S. experience = 15. *Case 3.* Age = 25–34, education ≥ 17, U.S. experience = 5.

sents the weight a woman places upon statistical evidence when making an estimate of total discrimination. This structural model of reporting discrimination yields the following econometric model:

$$(5)\quad \text{Prob}(\text{RD}_i = 1) = F\left[\frac{(\alpha b_j + c_j)X_{ij}}{\sigma_v(1-\rho^2)^{1/2}} - \frac{\rho\sigma_v\mu_i^F}{\sigma_F}\right],$$

where F is the standard normal cumulative density function; $c_j = \theta_j$ for $j \geq 1$, c_0 equals θ_0 minus a constant term; σ_F is the variance of μ_i^F; σ_v is the variance of the structural model's error term (v_i); and ρ is the correlation coefficient between μ_i^F and v_i.

$\Sigma_j \alpha b_j X_{ij}$ measures the contribution of statistical evidence to the probability of a woman reporting discrimination, and $\Sigma_j c_j X_{ij}$ measures the contribution of nonstatistical evidence. Kuhn correctly notes that α and σ_v are not identified. Hence, he normalizes $\alpha = \sigma_v = 1$. This is an innocuous normalization, but requires careful interpretation of parameter estimates.

The next step in Kuhn's analysis determines how much self-reported discrimination would decrease if all statistical discrimination were eliminated. The change in the probability of a woman's reporting discrimination if all statistical discrimination is eliminated is:[2]

$$(6)\quad F\left[\frac{\sum_j \hat{\tau}_j X_{ij}}{(1-\rho^2)^{1/2}}\right] - F\left[\frac{\sum_j \{\hat{\tau}_j - (\alpha/\sigma_v)\hat{b}_j\} X_{ij}}{(1-\hat{\rho}^2)^{1/2}}\right],$$

where $\hat{\tau}_j$ is the estimate of $(\alpha b_j + c_j)/\sigma_v$ obtained from the probit model's estimated coefficient for X_{ij}.

Since parameter estimates are available for b, τ, and ρ, it is evident from equation (6) that a value for the ratio (α/σ_v) must be assumed in order to compute the change in the probability of self-reported discrimination. Since α/σ_v is not identified, any prediction desired can be generated by making the appropriate assumption regarding (α/σ_v). Hence, Kuhn's simulation results hinge crucially upon the assumption that $\alpha = \sigma_v$. This assumption might be true, but there is no theoretical justification for assuming that it is true.

To illustrate the sensitivity of his results to the normalization chosen, consider the simulation results reported in Table 1. When $(\alpha/\sigma_v) = 1$, Kuhn's conclusions are generated.[3] For the group with median characteristics (case 1), the reduction in the probability of reporting discrimination in the United States is 5 percent when α/σ_v is 1, 9 percent when the ratio equals 2, and 14 percent when the ratio is 5. If α is negative (as found in his "naive" model), reported discrimination would actually increase. Hence, Kuhn's conclusions are erroneous unless, by chance, $\alpha = \sigma_v$.

All hopes of determining the importance of statistical discrimination should not necessarily be abandoned however. The task could be successfully accomplished if α/σ_v can be identified. One way to accomplish this would be to find a variable that is likely to affect *only* the statistical component of discrimination. If there is a theoretical basis for such an exclusion restriction (i.e., some X_{ij} should not be included as a component of ω_i), identification of α/σ_v is accomplished, and the appropriate simulations may be performed.

---

[2] In the simulation, the error term in the wage equation (μ_i) is assumed equal to zero, so the second part of equation (5) drops out of the analysis.

[3] In a conversation with Kuhn, he admits an error in calculating case 1 for the United States. The correct value (shown in Table 1) is 13.8 percent.

REFERENCE

Kuhn, Peter, "Sex Discrimination in Labor Markets: The Role of Statistical Evidence," *American Economic Review*, September 1987, *77*, 567–83.

Sex Discrimination in Labor Markets: The Role of Statistical Evidence: Reply

By PETER J. KUHN*

My 1987 paper offers three main contributions to an understanding of sex discrimination in labor markets: First, it establishes an apparently robust empirical result: confidential survey reports of discrimination are negatively correlated with standard statistical measures of wage discrimination. Second, it proposes an explanation of this result based on nonstatistical evidence of discrimination, and estimates a model incorporating it. Third, it attempts to draw some policy implications regarding the relative importance of nonstatistical evidence from the estimated parameters of that model.

In their comment on my paper, Debra Barbezat and James Hughes (BH) replicate my basic empirical analysis of the correlation between reports and measures using a substantially different data set. They also suggest an interesting alternative explanation of the results. William Even, in his comment, points out a conceptual error in the final, "policy implications" phase of my paper. In this Reply, I discuss the implications of both comments for the three components of my original paper in turn.

I. The Basic Results: How Similar Are BH's to Mine?

In my paper, I considered the correlation between statistically measured wage discrimination and answers to the question "Do you feel in any way discriminated against on your job because you are a woman?," in representative samples of U.S. and Canadian working women. In their paper, BH consider the correlation between statistically measured wage discrimination and answers to the question "Are you satisfied with the results of affirmative action at this institution?" in a sample of female college professors. These are, of course, potentially very different issues. College professors are a highly selected sample of individuals, compared to the general population. The measure of reported discrimination I focus on asks each woman whether, taking all forms of discrimination into account, she has herself been a victim of sex discrimination. BH's measure focuses on a particular *form* of discrimination (that involved in hiring and promotion), asks about discrimination *in the institution as a whole* rather than whether a woman herself feels she has been a victim, and does not limit attention to discrimination *on the basis of sex*, as opposed for example to race, which is also a target of affirmative action.

In view of these considerable differences, it is, to me, both encouraging and more than a little surprising that BH's basic results are so similar to mine, and apparently equally robust to changes in econometric technique.[1] I argue in the following section that this remarkable similarity might provide some clues to the reasons for the results.

II. Explaining the Basic Results: Nonstatistical Evidence Versus Quality of Statistical Evidence

Clearly, any "explanation" of the negative correlation between reports and statistical measures of discrimination will in some way be based on an omitted variable in the regression of reports on measures. Further, this omitted variable must be negatively correlated with measured discrimination and positively correlated with reports of discrimination. In my paper, I refer to this variable as "nonstatistical evidence," and think of it primarily as sex differences in nonwage compensation and treatment on the job that are observed by the worker but not the econometrician. This was motivated primarily by the way the discrimination question was phrased—"were you discriminated against *in any way* because you are a woman?"—combined with a casual sense that discrimination can take many forms besides the kind of wage discrimination economists typically measure. More generally, however, it is clear that nonstatistical evidence is simply a residual category which comprises all "evidence of discrimination that is not captured by standard wage regression techniques": (p. 568), including any "other information about wage discrimination not embodied in D_i" (p. 574).

---

*Department of Economics, McMaster University, Hamilton, Ontario, Canada L8S 4M8. Some of the work on this Reply was done while the author was visiting the Department of Economics and the Industrial Relations Section at Princeton University. I would like to thank (without implicating) David Card, William Even, Claudia Goldin, Steven Jones, and Mike Veall for helpful comments.

[1] Some of the robustness tests I conducted, such as excluding education and experience from the discrimination measure, or correcting for sample selection bias, are not conducted or are not feasible in BH's study.

Aside from the basic requirements listed above (negative correlation with measures, positive with reports), can we learn anything else about the omitted variable we call "nonstatistical evidence"? One approach to this problem is the one I take in my formal model of reporting behavior. In that model I am able to estimate which observed features of the data are associated with high levels of nonstatistical evidence. A clear answer emerges: young, well-educated women seem to face more of this "evidence" than others. This allows me to speculate somewhat further on its possible nature (p. 529): perhaps young women are still searching for low-discrimination firms; perhaps more-educated women are more perceptive of readily available evidence; or perhaps more-educated women have better estimates of wage discrimination because they are more likely to have males in similar jobs as a comparison group.[2] (Note that this last possibility is very closely related to BH's hypothesis.) Finally it is also possible that attitudes and expectations of women have been changing over the years and are affected by education. As a result, younger, more educated women may simply be more willing to label as discrimination and less willing to tolerate behavior considered acceptable, for example, by their mothers.[3]

A second approach to learning more about the "missing variable" is to find ways to measure and control for it in the regression of reports on measures. If the negative correlation weakens or disappears, the variable in question is potentially part of the explanation. This is precisely what BH do, in a particularly ingenious test that uses institutional type as a proxy for the quality of statistical evidence (their proposed "missing variable"). Unfortunately, their test results are mixed. It thus appears that quality of statistical evidence tells at best only part of the story. Indeed, given the reluctance of the basic negative correlation to disappear across data sets, definitions of discrimination, and (as BH now find) types of firms, I am now somewhat more inclined toward the notion of changing general attitudes with age and education as a most likely explanation. A lot more work on the issue however needs to be done; in the remainder of this section I present some suggestions. Much of it involves some questions that can easily be answered using the BH data itself.

One aspect of BH's data that remains unexplored is the question of *which* women are most likely to report discrimination. In particular, is the impact of age on reporting negative, as in my study? And even though there is likely little variation in education in their sample, do it and related factors (like academic rank, for example) raise the probability of reporting as I found? If age and education have similar effects here to those in my study, our conclusions about the nature of the "missing variable" determining reports of discrimination become somewhat stronger.

---

[2] Yet another interpretation emerges from simple models of occupational segregation based on firm-specific training (for example, Kuhn, 1988), in the following way: Suppose reports of discrimination depend on current male-female *productivity* differentials as well as wage differentials (lower relative productivity of males, holding wages fixed, contributing to a greater perception of discrimination). Then to the extent that young women are assigned to jobs that require less on-the-job training (and thus have higher *current* productivity), current wage differentials will understate the "true" discrimination that they face.

[3] Direct evidence of such changes in general attitudes is found in Thornton et al. (1983).

Related to the above point, it would be useful to know whether the overall percentage of women reporting discrimination is higher or lower in BH's sample of college professors than in my sample of the general population.[4] If higher, this would provide further support for the notion that something about high levels of education is conducive to reporting discrimination, even when statistical evidence is controlled for, as I found. Finally, it would be interesting to conduct direct tests of some of BH's maintained hypotheses against my own. For example, I assume that women, because they observe nonstatistical evidence *in addition to* statistical evidence, know *more* than the econometrician about whether they, personally, are victims of discrimination. BH, on the other hand, assume women know *less*. A survey in which women were asked how much pure wage discrimination they believed they faced in a clearly phrased question might help resolve this question.

III. Relative Importance of Statistical and Nonstatistical Evidence: Corrected Estimates

In response to a demand for policy implications of my results, I attempted in the last phase of my original analysis to produce a summary estimate of the relative importance of statistical and nonstatistical evidence that might have some bearing on the effects of recent Supreme Court decisions. The quantitative results I offered on this point were estimates of the percentage of women who would continue to report discrimination even if all statistical evidence of wage discrimination were ignored or eliminated.

In his comment, William Even correctly points out that the estimates I presented depend on my normalizing assumption that $\alpha/\sigma_v = 1$. I had concluded, erroneously, that this was not the case. Even's comment thus forces me to reconsider the meaning of "relative importance" in Part A below. I show there that there are at least two very distinct meanings of this term, and that it was my failure to distinguish clearly between them that led to the mistake Even identifies. Since both notions of relative importance can still be estimated empirically Parts B and C below develop and present these corrected empirical results in turn.

A. *What Do We Mean by "Relative Importance"?*

The probability of reporting discrimination in my model is given by (5) in Even (1988). Suppose for a moment that all women face exactly the *same* amount of statistically measured wage discrimination, 0.3 in log-wage units, but that they experience *different* amounts of nonstatistical evidence, with a mean of 0.1. Further assume they weigh both types of evidence in their reporting decisions equally ($\alpha = 1$). In this example, the proportion of the variance in total evidence of discrimination (and hence in reports of discrimination) explained by statistical evidence is zero. Thus in a sense, statistical evidence is not an important determinant of women's reporting decisions. In another sense, however, statistical evidence is more important: it is three times larger on average, and its elimination might thus cause a large reduction in the fraction of women who claim to have experienced discrimination.

There are, then, at least two meanings of the term "relative importance." The first reveals whether differences in statistical or nonstatistical evidence determine why some women report discrimination while others do not. Provided $\alpha > 0$, conclusions about this concept can in some cases be drawn even if α/σ_v is not exactly identified, as is demonstrated formally in Part B below.[5] The second notion of relative importance refers to the average levels of the two types of evidence. It is essentially this measure of relative importance that must be identified to estimate the effects of removing one type of evidence, and as Even points out, this requires knowing α/σ_v.[6] To the extent that U.S. courts will weight statistical and nonstatistical evidence the same way individuals do, *both* these measures have interesting policy implications. For example, the first, "variation" measure contains information on how much better the courts would be able to distinguish women who are discriminated against from those who are not, when statistical evidence, as opposed to personal commentary, is introduced. The "levels" measure tells us, on the other hand, by how much total "effective" evidence of discrimination will increase, on average, if statistical evidence is allowed in court.[7] For this reason I present estimates of both types of relative importance below.

---

[4] BH indicate in a footnote that the levels of statistically measured discrimination are 0.065 and 0.135 for public and private institutions respectively. This is considerably below the levels of 0.207 and 0.396 I found in Canada and the United States, respectively, perhaps because of their superior human capital controls. They apparently do not report the overall means of the reported discrimination dummy, RD_i.

B. *Estimating "Relative Importance"*

A natural measure of the relative importance of statistical versus nonstatistical evidence in determining the overall cross-sectional variation in reports is given by:

$$(1)\quad \frac{\mathrm{Var(Stat)}}{\mathrm{Var(NonStat)}} \doteq \frac{\mathrm{Var}\left(\sum_j \alpha b_j X_{ij}\right)}{\mathrm{Var}\left(\sum_j \theta_j X_{ij}\right)} = \frac{\mathrm{Var}\left(\sum_j \alpha b_j X_{ij}\right)}{\mathrm{Var}\left(\sum_j c_j X_{ij}\right)}.$$

This is just the ratio of the (systematic) variance of effective statistical evidence to nonstatistical evidence of discrimination across all women in the sample. Equation (1) can be rewritten, dropping subscripts, as:

$$(2)\quad \frac{\mathrm{Var(Stat)}}{\mathrm{Var(NonStat)}} = \mathrm{Var}(\Sigma bX)/\{\mathrm{Var}(\Sigma bX) + \lambda[\lambda \mathrm{Var}(\Sigma \tau X) - 2\,\mathrm{Cov}(\Sigma bX, \Sigma \tau X)]\}$$

where $\lambda = \sigma_v/\alpha$ and $\tau_j = (\alpha b_j + c_j)/\sigma_v$. This expression is of course zero when $\mathrm{Var}\,\Sigma bX = 0$, and infinite when $\mathrm{Var}\,\Sigma cX = 0$. Given $\alpha > 0$, a sufficient condition for this ratio to be less than one (and therefore, for statistical evidence to be, in a well-defined sense, a less important determinant of individual reports of discrimination) is that:

$$(3)\quad \mathrm{Cov}(\Sigma bX, \Sigma \tau X) < 0.$$

In words, the covariance of statistical evidence and estimated "total evidence" of discrimination must be negative. This is easily checked using coefficients from the original model and does not require knowledge of α/σ_v.

---

[5] The intuition is apparent from the following example. Consider the difference, in the United States, between a woman with less than 12 years of education and one with more than 17 years. From Table 5 of the original paper, the woman with more education faces less statistical evidence ($b_j = -0.204$), yet is more likely to report discrimination ($\tau_j = 1.342$, using Even's notation). Now, using $c_j = \sigma_v \tau_j - \alpha b_j$, as long as $\alpha > 0$ (i.e., assuming women place some positive weight on statistical evidence) it must be the case that $c_j > 0$ and $|c_j| > |\alpha b_j|$. Thus the effect of an increase in years of education on the level of nonstatistical evidence must outweigh its effect on "effective" statistical evidence to generate the observed pattern of reports.

[6] Actually, the total amount of nonstatistical evidence $\Sigma_{j=0}^{k} \theta_j X_{ij}$, need not be identified to estimate the effect of removing statistical evidence. What needs to be known is the amount of statistical evidence, $\Sigma_{j=0}^{k} b_j X_{ij}$, and nonstatistical evidence *net* of the mean reporting threshold, $\Sigma_{j=0}^{k} c_j X_{ij}$.

[7] It is important to emphasize, however, that these policy conclusions can be drawn only if we assume that individuals observe the same measures of statistical and nonstatistical evidence that the courts eventually observe, and that the unobserved variable, "nonstatistical evidence," takes the form of discriminatory treatment not seen by the econometrician. In particular, to the extent the unobserved variable represents a change in attitudes or expectations such as those discussed above, these policy conclusions cannot be drawn.

The second measure of relative importance, unlike the first, cannot be estimated unambiguously in the original model. It can, however (as Even suggests), be identified with the aid of exclusion restrictions. I follow that suggestion here by asking the following question: Suppose that none of the independent variables in Table 5 of the original paper, except for those the model indicates are clearly important (namely the age and education dummies), had any effect on the level of nonstatistical evidence, ω_i.[8] What then is the maximum likelihood estimate of α/σ_v implied by each of the two data sets in question? And what in turn are the model's predictions about the effects of removing all statistical evidence of discrimination?[9]

Under the above restrictions, the probability of reporting discrimination for an individual woman can be written:

$$(4)\quad \mathrm{Prob}(\mathrm{RD}_i=1) = F\left\{\frac{\delta_0 D_i^2 + \sum_{j=1}^{9}\delta_j X_{ij} - \rho\tilde{\mu}_i^F}{\sqrt{1-\rho^2}}\right\},$$

where $\delta_0=\alpha/\sigma_v$; $\delta_j=c_j/\sigma_v$; D_i^2 is statistically measured wage discrimination

$$\left(\sum_{j=1}^{k} b_j X_{ij}\right); \qquad X_{i1}-X_{i9}$$

are the age and education dummies plus a constant term; $\tilde{\mu}_i^F$ is the standardized error term in the female earnings equation (μ_i^F/σ_F); ρ is the correlation coefficient between μ_i^F and v_i, the error term in the probit for reporting; and F is the standard normal distribution function. Replacing the corresponding terms in Table 4 of the original paper by (4) yields a log likelihood function which can be maximized using standard techniques; all parameters in this new function (in particular $\delta_0=\alpha/\sigma_v$) are identified without normalizing assumptions.

C. *Results*

The simple correlation coefficient between the level of statistical evidence $(\Sigma_j b_j X_{ij})$ and the propensity to report discrimination $(\Sigma_j \tau_j X_{ij})$ in the unrestricted model of reporting (using the coefficients in Table 5 of the original paper) is -0.069 in Canada and -0.349 in the United States. This correlation is significantly different from zero at the 99 percent level in the United States ($t=-7.17$) but not so in Canada ($t=-1.45$).[10] The sufficient condition (3), for nonstatistical evidence to be a more important determinant of the propensity to report discrimination is thus satisfied by the point estimates in both data sets. The "variation" measure of relative importance thus bears out a key intuition from the original paper, independent of α/σ_v: it is only because it *outweighs* the effects of statistical evidence which work in the opposite direction that cross-sectional variation in the level of nonstatistical evidence can explain the pattern of reports seen in these two data sets.

---

[8] We can conclude that age and education are important in the unrestricted model because they both have strong effects on reporting (τ_j), which are opposite in direction to their effects on statistical evidence (b_j).

[9] Admittedly, these exclusion restrictions have little theoretical appeal. However, since they are more than what is strictly necessary to identify the model (setting only one of the c_j's equal to zero would be sufficient to identify α/σ_v), they can be evaluated empirically. In particular, the restricted models involve 8 overidentifying exclusion restrictions in Canada and 9 in the United States. In neither case can these be rejected at any conventional level of significance. As well, experiments with different sets of excluded variables using less cumbersome two-stage estimation techniques indicate the results are not very sensitive to the precise set of restrictions used, with one exception. In particular, for Canada, the two-stage estimate of α/σ_v in the present model was 1.98. When the restriction $\rho=0$ was added, this changed to 1.96. When the restriction $c_j=0$ for the trade school, union, and physical disability dummies was removed, it changed to 1.68. Corresponding numbers for the United States are 0.68, 0.67, and -0.18.

[10] The t-statistic for the test of $\rho=0$ was calculated from Hughes and Grawoig (1971, p. 349). This test ignores the fact that b_j and τ_j are estimated coefficients, and should therefore be interpreted with caution.

TABLE 1—FIML ESTIMATES OF STRUCTURAL MODEL PARAMETERS: AGE AND EDUCATION VARIABLES

		Canada		United States	
		b_j	δ_j	b_j	δ_j
Intercept		0.177*	−2.348*	0.230*	−2.478*
		(0.083)	(0.486)	(0.115)	(0.572)
Age	25–34	−0.017	0.292	−0.001	0.460
		(0.085)	(0.341)	(0.091)	(0.308)
	35–44	0.083	−0.418	0.018	0.143
		(0.092)	(0.374)	(0.118)	(0.340)
	45–54	0.155	−0.330	0.125	−0.675
		(0.103)	(0.458)	(0.139)	(0.431)
	55–76	0.144	−1.139	0.045	−0.169
		(0.128)	(0.693)	(0.164)	(0.477)
Education:	12 years	−0.008	0.641*	−0.041	0.874*
		(0.069)	(0.326)	(0.084)	(0.425)
	13–15 years	0.046	0.849*	0.021	0.531
		(0.086)	(0.394)	(0.089)	(0.461)
	16 years	−0.063	1.428*	0.089	0.405
		(0.114)	(0.547)	(0.104)	(0.499)
	17+years	−0.065	1.773*	−0.211*	1.653*
		(0.125)	(0.578)	(0.102)	(0.530)
Estimate of	$\delta_0 = \alpha/\sigma_v$	3.174*		1.216	
		(1.593)		(0.890)	
Number of Observations		1106		992	
Log Likelihood		−703.95		−537.01	

*Denotes a coefficient significantly different from zero at the 95 percent level.

Note: The b_j coefficients measure the effect of characteristic X_j on the level of statistically measured wage discrimination. The δ_j coefficients for age and education give the (standardized) effects of these variables on the level of nonstatistical evidence, c_j/σ_v. Standard errors are in parentheses.

FIML estimates of b_j and δ_j for the age and education variables under the proposed exclusion restrictions are given in Table 1. The parameter of greatest interest is the estimate of $\delta_0 = \alpha/\sigma_v$, which equals 3.174 in Canada and 1.216 in the United States. Since nothing in the estimation techniques constrains δ_0 to be positive, these results are encouraging, and suggest that women do place positive weight on statistical evidence in their reporting decision, as is maintained here. The parameter magnitudes also appear plausible, although the estimate for Canada is considerably higher than the value of 1 assumed in the original paper. Estimates of the other coefficients exhibit the same pattern as in the original paper: nonstatistical evidence decreases with age and increases with education, while the opposite is true for statistical evidence.

Simulations of the effect of removing all statistical evidence on reporting in the new model are given in Table 2 which, in addition to the three cases examined in the earlier paper and in Even (1989), considers what would happen to the entire sample of women. In all comparable cases, the reduction in reports when statistical evidence is removed is greater than in the original paper, reflecting in part the estimates of α/σ_v in excess of 1. Statistical evidence is thus more important than originally estimated.[11] In both countries, however, over 35 percent of young, well-educated women, and over 6 percent of all women in the sample would continue to report discrimination in the absence of sta-

---

[11] This is especially true for women with median characteristics in Canada, who experience very large declines in reporting if statistical evidence is removed. If the Courts' relative weights and standards of proof are similar to individuals', these women could gain considerably if statistical evidence were ruled admissible in Court.

TABLE 2—PREDICTED PROBABILITIES OF REPORTING DISCRIMINATION

	(1) Using All Evidence	(2) Using Nonstatistical Evidence Only	(3) Change (1)–(2)
Canada:			
Case 1 (Median)	0.1220	0.0161	0.1059
Case 2 (Older, Less Educated)	0.0102	0.0002	0.0100
Case 3 (Younger, More Educated)	0.5744	0.3876	0.1868
Case 4: Sample Average	0.1549	0.0625	0.0924
United States:			
Case 1 (Median)	0.1697	0.0704	0.0993
Case 2 (Older, Less Educated)	0.0221	0.0038	0.0183
Case 3 (Younger, More Educated)	0.4331	0.3564	0.0767
Case 4: Sample Average	0.1317	0.0680	0.0637

Notes: Case 1: Age 35–44, 12 years school; 10 years total experience in U.S. only.
Case 2: Age 55–76, <12 years school; 15 years total experience in U.S. only.
Case 3: Age 25–34, ≥ 17 years school; 5 years total experience in U.S. only.

Cases 1–3 all consider nonunion workers, with no trade school or physical disability, 2–3 years tenure, and $\mu_i^F = 0$; previous experience dummy $=1$ in Canada only.

Case 4 calculates a predicted report probability for each woman in the sample as a function of all characteristics including $\hat{\mu}_i^F$ (her earnings function residual), and then averages these over the entire sample.

tistical evidence. In Canada, this 6 percent represents about 4 out of every 10 women who report experiencing discrimination based on all available evidence; in the United States it is 5 in 10. Nonstatistical evidence thus remains substantial, especially for young, well-educated women. Also, as noted earlier, because more-educated women experience less statistically measured discrimination, nonstatistical evidence remains the only possible explanation for their high reporting rates in the context of the present model.

IV. Summary

The comments by BH and Even both contribute substantially to our understanding of the role of different kinds of evidence in women's perceptions of sex discrimination. BH replicate my basic empirical analysis of the correlation between reported and measured discrimination on a different data set, suggest an alternative interpretation of the results, and test that interpretation against the data. The mixed results of this test suggest that their "quality of statistical evidence" hypothesis tells at best only part of the story behind the observed negative correlation; hopefully further work will help sort out a number of other alternative explanations.

Even points out an error in the interpretation of my original results for policy purposes, which I have corrected in this note. The corrected analysis shows the following. First, provided that women place *some* positive weight on statistical evidence, we can conclude that nonstatistical evidence explains a larger fraction of the observed *variation* in women's reports of discrimination than does statistical evidence. This notion was implicit in the discussion of results in my original paper, but never formally demonstrated.

Second, the removal of statistical evidence of discrimination is predicted to reduce reports of discrimination in a representative sample of full-time working women from 15.5 percent to 6.3 percent in Canada and from 13.2 percent to 6.8 percent in the United States. These declines are larger than those predicted in my original paper and depend on exclusion restrictions not imposed there. Still, at least 4 in 10 women who report discrimination in both data sets are predicted to *continue* reporting discrimination in the absence of any statistical evidence.

In sum, the corrected estimates presented in this Reply show that, in terms of both size and variance, nonstatistical evidence of discrimination remains substantial. At some level, this should hardly be surprising: Discrimination can, after all, take many forms besides sex-based wage differentials, and seems unlikely to be completely captured by so crude a measure as a log-earnings regression. Perhaps the present analysis will stimulate further research on these other, relatively neglected aspects of sex discrimination.

REFERENCES

Barbezat, Debra A., "Affirmative Action in Higher Education: Have Two Decades Altered Salary Differentials by Sex and Race?" *Research in Labor Economics*, 1989, *10*, 107–56.

Even, William, "Sex Discrimination in Labor Markets: Comment," *American Economic Review*, March 1990, *80*, 287–89.

Hughes, Ann and Grawoig, Dennis, *Statistics: A Foundation for Analysis*, Reading, MA: Addison-Wesley, 1971.

Kuhn, Peter. "Sex Discrimination in Labor Markets: The Role of Statistical Evidence," *American Economic Review*, September 1987, *77*, 567–83.

______, "A Simple Model of Occupational Segregation," McMaster University, mimeo., November 1988.

Thornton, Arland, Alwin, Duane F. and Camburn, Donald, "Causes and Consequences of Sex-Role Attitudes and Attitude Change," *American Sociological Review*, April 1983, *48*, 211–27.

Self-Selection and the Earnings of Immigrants: Comment

By GUILLERMINA JASSO AND MARK R. ROSENZWEIG*

In a recent article in this *Review* (Borjas, 1987), George Borjas uses a standard model of self-selection to demonstrate that immigrants in the United States may not necessarily be positively selected, as is commonly assumed. Borjas' empirical results, based on U.S. Census micro data merged with country-of-origin characteristics for 41 sending countries, however, do not provide clear or consistent results with respect to the immigration selection issue, although they do confirm earlier findings that the characteristics of the countries of origin of U.S. immigrants explain a substantial proportion of the differences in their economic status (Jasso and Rosenzweig, 1986b).[1] The purpose of this note is to show that Borjas' results are themselves subject to biases due to additional processes of "self-selection." In particular, his inconclusive findings with respect to immigration selectivity arise from the use of a choice-based sample; enlargement of his highly selective sample of countries, based on the number of foreign-born in the United States, yields results that appear to conform much more closely and consistently to conventional wisdom, namely that the United States appears to attract those persons with above-average skills. We also provide evidence that changes in the earnings of aggregate immigrant entry cohorts, used by Borjas to measure the "assimilation" of immigrants, reflect another evidently important selection process, namely selective emigration. The effects of country characteristics on the wages of the foreign-born at entry (immigration selectivity) are thus not comparable to the effects of such variables on the changes in the characteristics of an immigrant cohort over time (emigration selectivity).

The effects via immigration self-selection of country-of-origin characteristics on the "quality" Q of immigrants who migrate to a destination country is neatly summarized by Borjas in his equation (12). With unobservables determining earnings normally distributed, the expected quality of immigrants in the United States is given by

$$E(Q)=\mu+\gamma\lambda, \tag{1}$$

where $\mu =$ earnings of the U.S. native born and $\lambda=\phi(z)/P$, where P is the emigration rate of the origin country and $\phi(z)$ is the density of the standard normal. Immigrants are positively selected if $\gamma>0$ and negatively selected if $\gamma<0$. The effect of, say, migration costs on $E(Q)$ is thus

$$\frac{\partial E(Q)}{\partial c}=\gamma\frac{\partial\lambda}{\partial z}\frac{\partial z}{\partial c}. \tag{2}$$

Since $\partial\lambda/\partial z>0$, and $\partial z/\partial c>0$ (as migration costs increase fewer persons migrate), increases in migration costs raise (lower) immigrant quality when immigrants are positively (negatively) selected. Similarly, when mean incomes (earnings opportunities) rise in the origin country, emigration declines and the quality of immigrants rises or falls depending on whether selectivity is positive or negative. Whatever the type of selectivity, mean origin-country income and migration costs should affect the quality of immigrants in the same way.

Borjas' empirical results with respect to origin-country mean income and migration costs are inconsistent with the model. While he finds that distance—the measure of migration costs—and mean origin-country income reduce emigration rates, mean origin-country income is positively associated with the initial wages of immigrants in the United States and with his assimilation measure, while distance has no effect on entry wages

---

*Department of Sociology, University of Iowa and Department of Economics, University of Minnesota, Minneapolis, MN 55455. Partial support for this research was provided by the Russell Sage Foundation.

[1] See also Greenwood and McDowell (1986).

TABLE 1—MEAN CHARACTERISTICS OF COUNTRIES IN BORJAS SAMPLE AND ADDITIONAL SAMPLE COUNTRIES

	Borjas Sample ($N=41$) (1)	Additional Countries ($N=66$) (2)	Extended Sample ($N=107$) (1)+(2)
Immigrants, 1975–80	55240	9395	27127
Foreign-Born Males Aged 25–64, Entered 1975–80[a]	600.3	133.2	312.2
Population Size (1970), in Thousands	41831	23122	30291
Distance, Miles	3745	5273	4687
Per Capita, GNP (1978), Dollars	3810	1775	2555
Centrally Planned	0.146	0.0606	0.0934
Literacy Rate (1975), Percent	84.8	67.8	74.3
Western Hemisphere	0.317	0.273	0.290
Proportion of World Population (1970)	0.498	0.421	0.919
Proportion of U.S. Foreign-Born Entered 1975–80 in 1980	0.661	0.226	0.887

[a] From *1980 Public Use Tapes*, 2.5:100 household A sample.

and has a negative or zero effect on assimilation. Borjas selects his sample of countries, however, on the basis of the number of foreign-born from each country residing in the United States in 1970; those countries with numbers of foreign-born below a minimum are selected out in order to maintain a minimum per country sample size, a constraint that is not relevant to the analysis of the initial earnings of the recent-entrant foreign-born in 1980. If this sample selection rule not only excludes those countries with lower absolute numbers of migrants in the United States but also those with lower emigration *rates*, then the effects of country-of-origin characteristics on emigration rates will be biased (to zero) by this sample truncation. And, from (2), if dz/dc, say, is biased to zero, then so will the estimated selectivity effects of changes in c.

In Table 1 we present the mean characteristics of the 41 countries selected by Borjas and an additional 66 countries for which we could obtain comparable characteristics (and used by us in prior work (Jasso and Rosenzweig, 1986a, 1986b, and forthcoming)).[2] As can be seen, Borjas' country sample accounts for less than half of the world's population and only 66 percent of the foreign-born who entered the United States between 1975 and 1980. The addition of the 66 countries extends the coverage to 89 percent of the world's population and to 92 percent of the newly entered U.S. foreign-born in 1980. Most importantly, the excluded countries (accounting for 23 percent of the foreign-born) have a substantially lower average emigration rate in the 1975–80 period (0.040 versus 0.132 percent), are located significantly farther from the United States on average (5273 versus 3745 miles), and are characterized by lower per capita GNP and literacy rates.

In Table 2 we report grouped probit (minimum-chi square) regressions of the emigration rate based on the Borjas and ex-

---

[2] Among the countries excluded from the Borjas sample but included in the extended sample are 10 countries in Africa, including South Africa, Liberia, Nigeria, and Kenya; 11 countries in South and Central America, including Chile, Costa Rica, El Salvador and Nicaragua; and Turkey, Syria, mainland China, Indonesia, Pakistan, Malaysia, Thailand, Vietnam, Australia and New Zealand. Sources of country characteristics are provided in Jasso and Rosenzweig (1986a).

TABLE 2—PROBIT REGRESSIONS OF EMIGRATION RATE BASED ON U.S. FOREIGN-BORN MEN AGED 25–64 ENTERED 1975–80[a]

Country of Origin Characteristics	Borjas Sample (1)	Borjas Sample (2)	Extended Sample (1)	Extended Sample (2)
Per Capita GNP ($\times 10^{-4}$)	0.0672	0.0160	−0.0328	−0.556
	(0.94)[b]	(0.24)	(0.31)	(3.43)
Distance ($\times 10^{-4}$)	−0.212	0.0481	−0.537	−0.776
	(1.77)	(0.35)	(9.54)	(9.91)
Literacy Rate	–	0.00234	–	0.00245
		(3.10)		(4.10)
Not Centrally Planned	−0.00793	0.0341	−0.116	−1.52
	(0.34)	(0.18)	(11.8)	(12.0)
Western Hemisphere	−0.221	−0.200	−0.211	−0.325
	(6.22)	(6.14)	(6.06)	(7.61)
Constant	−4.37	−4.70	−4.11	−3.96
	(53.2)	(36.0)	(69.2)	(58.7)
Number of Countries	41	41	107	107
χ^2	156.1	156.6	56.1	55.3

[a] Denominator is country-of-origin population in 1970.
[b] *t*-ratios in parentheses.

tended country samples, where we use the number of foreign-born residing in the United States in 1980 who entered in the period 1975–80 divided by origin-country population size in 1970 as the dependent variable. The number of resident foreign-born, based on 1980 Census data, is used to measure emigration because the number of legal immigrants, used by Borjas in his emigration regressions, does not correspond to the foreign-born population in the Census from which Borjas estimates his wage functions. The latter population includes, particularly among the newly entered foreign-born, persons with temporary visas permitting employment, foreign students, and various types of persons who are "illegal" immigrants. The resident foreign-born and legal immigrant populations are quite different—for example, of the foreign-born in 1980 who entered between 1975 and 1980, 20.2 percent are from Mexico, while only 14.9 percent of the total number of (legal) immigrants came from Mexico in the same period. Our emigration estimates thus more closely correspond to the selection equation underlying Borjas' entry wage function estimates.[3]

Results from two specifications are reported in Table 2, one with and one without the country's literacy rate (excluded by Borjas).[4] With either specification, we see that the use of the truncated country sample results in a severe underestimate of the negative effect of distance on the probability of immigration to the United States, by a factor of over two with literacy rates excluded. When literacy rates are included, neither the effects of per capita GNP nor of distance are statistically significant in the Borjas sample; in contrast, both variables have the expected negative sign and are highly significant in the extended sample. We interpret the results in the last column of Table 2, estimated on the extended sample, as indicating that the United States offers higher returns, on average, to skills compared to other countries. Origin-countries with the same per capita income, but higher literacy rates, are characterized by lower returns to skills; such countries have significantly higher emigration rates. Similarly, countries with the same literacy rates but higher levels of per capita income are likely to have higher returns to skills, and exhibit significantly lower rates of emigration to the United States.

---

[3] There are other reasons why the use of the Immigration and Naturalization Service (INS) legal immigration data is less desirable in this case. First, prior to 1982, the INS data do not distinguish between immigrants from mainland China and Taiwan, two countries with a distinctly different set of characteristics. While it is possible to identify the foreign-born from Taiwan (one of Borjas' set of 41) and mainland China based on the Census data, the "Chinese" immigrants used in Borjas' emigration regressions based on the INS data are from both mainland China and Taiwan. The INS immigration data used by Borjas also include both men and women (and children), while his wage analysis examines only adult men. In our earlier work (Jasso and Rosenzweig, 1986a and forthcoming) we have found the immigration behavior of men and women to be distinctly different; a larger proportion of legal immigrant women compared to men enter the United States via marriage to a U.S. citizen, while a larger percentage of men compared to women enter by dint of their labor market skills. As a consequence, for example, the presence of a U.S. military base in an origin country importantly influences the flow of female but not male immigrants to the United States from that country.

[4] Borjas' list of country characteristics also included two "political" variables and a measure of income inequality. These variables are not available for most of the countries in the extended sample. One political variable and income inequality did have statistically significant effects on emigration. Thus our emigration results may be subject to specification error, but it is not

Does sample truncation also account for the absence of an effect of distance on the "entry" wage of the foreign-born in Borjas' results? In Table 3 we present estimates of the determinants of the log of the hourly wage rate of foreign-born males aged 25–64 who entered the United States in 1975–1980. We selected the same sample as Borjas and use the same Census-based regressors.[5] Our specification differs from his in that he allowed the effects of personal characteristics on the wage to differ across countries but assumed the effects of these variables to be the same within countries for different entry cohorts. We look at only one entry cohort, but assume that the effects of personal characteristics do not differ across countries.[6] Results based on the Borjas sample (41 countries) are reported in column 1. Despite our somewhat different specification, we are able to replicate the insignificant effect of distance and the positive and significant effect of per capita GNP on the entry wage rate with the truncated Borjas sample.[7]

In the second column of Table 3, we report the entry wage function estimates using the extended sample. While the effects of the personal characteristics on the wage are not sensitive to country selection, distance becomes statistically significant (at the 0.05 level) and its magnitude almost doubles when the sample is not truncated. The positive effects of both per capita GNP and of distance on the wage rates of the foreign-born are consistent with positive immigrant selection with respect to unobservables that influence wages, given that both variables inhibit immigration to the United States. The marginally significant literacy rate effect may be due to literacy measuring, in part, schooling quality; the sign of its effect does not therefore constitute as clear a test of selectivity as do the signs of the distance and origin-country income coefficients.

Borjas' additional findings that distance significantly reduces or has no effect on "as-

---

obvious how such specification error is responsible for the changes in estimates across the samples. Moreover, none of the excluded variables were found by Borjas to affect entry wages or assimilation rates. The only skill variable included by Borjas in the emigration equation was a measure of the English proficiency of the U.S. residents from the origin country, which we believe to be a poor measure of the average English proficiency, or average skill level, of the total *origin* population, given immigrant selectivity. We include, as did he, a dummy variable indicating whether or not the country is in the Western Hemisphere to capture, in part, the effects of the differential immigration criteria existing for the two hemispheres in the period between 1965 and 1978.

[5] The sample consists of employed (in 1979) foreign-born men aged 25–64 who were not self-employed or working without pay in 1979 and who were not in the Armed Forces or a resident of group quarters during the survey week. The sample was drawn from the 1:100 *A Sample of the 1980 Census Public Use Tape*.

[6] Borjas estimated separate wage functions based on Census samples from each of the 41 countries including a random sample of comparable U.S. native-born residents. It appears that the specification used assumed that the effects of all personal characteristics (see Table 3) except years in the United States were identical for the native-born and foreign-born populations. Since the reference population (U.S. native-born residents) is the same for all immigrants, we only employ a sample of the U.S. foreign-born.

[7] The variables excluded from our specifications, including income inequality, a variable cogently highlighted by Borjas in his theoretical analysis, did not have a statistically significant effect on either entry wage differentials or assimilation rates in Borjas' specification.

TABLE 3—DETERMINANTS OF LOG OF HOURLY WAGE RATE OF NEW ENTRANT FOREIGN-BORN MEN AGED 25–64 IN 1980

	Borjas Sample	Extended Sample
	(N = 1949, Countries = 41)	(N = 2534, Countries = 91)
Country Characteristics		
Per Capita GNP ($\times 10^{-4}$)	0.699	0.664
	(7.29)[a]	(8.20)
Distance ($\times 10^{-4}$)	0.123	0.242
	(0.76)	(2.21)
Literacy Rate ($\times 10^{-3}$)	0.510	1.18
	(0.44)	(1.37)
Not Centrally Planned	0.229	0.0998
	(3.29)	(2.14)
Western Hemisphere	−0.0696	0.00363
	(0.75)	(0.05)
Individual Characteristics		
No English or English Not Well	−0.202	−0.198
	(4.62)	(5.22)
Age	0.107	0.0919
	(7.34)	(7.16)
Age Squared	−0.00122	−0.00105
	(6.84)	(6.73)
Schooling Attainment	0.0206	0.0231
	(4.75)	(6.31)
Married, Spouse Present	0.109	0.133
	(2.69)	(3.73)
SMSA	0.197	0.133
	(2.62)	(3.73)
Disability	−0.101	−0.142
	(0.79)	(1.26)
Constant	−1.16	−0.770
	(2.74)	(12.13)
R^2	0.211	0.190
F	43.2	49.2

[a] t-ratios in parentheses.

similation" while origin-country GNP increases it would appear to contradict our conclusions about positive immigration selectivity (although they are not consistent with negative selectivity either). However, aside from sample truncation, the change in the earnings of a foreign-born cohort defined by date of and age at entry to the United States, used by Borjas to measure assimilation, does not necessarily correspond to the change in earnings for a randomly chosen "typical" immigrant.[8] The survivors of an age-entry cohort after a period of time has elapsed are not the same people as the original entrants due to deaths and, most importantly, to the emigration or re-migration of the foreign-born.[9] Distance may impede emigration from the United States, as it does to the United States. If foreign-born emigrants (from the United States) are negatively selected (for example, returnees are those who fare less well than they expected) then the progress of an aggregate immigrant cohort may very well be negatively associated with distance even though immigrant selectivity is positive.

---

[8] Borjas does not actually track or compare each cohort over time, as he pools men of the same age in both 1970 and 1980. However, as he notes, this procedure is equivalent if the appropriate parameters describing the relations of interest are stable over time.

[9] Another reason is a change in the coverage of the Census. If the coverage change is selective by country-of-origin and income, inferences about differential assimilation rates (or the selectivity of emigration) based on aggregate cohort data from adjacent censuses are made even more difficult.

TABLE 4—DECADAL COHORT CHANGE (PERCENT) IN RATIO OF FEMALES TO MALES BY AGE GROUP AND NATIVITY: 1960–70 AND 1970–80

	Cohort Age in 1960			Cohort Age in 1970		
Nativity	20–29	30–39	20–44	20–29	30–39	20–44
Native Born	−0.48	2.90	0.97	−2.90	2.53	−0.97
New Entrants[a] —Eastern Hemisphere	23.9	1.14	13.0	49.7	9.38	24.0
New Entrants—Western Hemisphere	68.4	21.3	43.4	5.60	5.86	−2.36

Source: 1960, 1970, and *1980 Public Use Tapes.*

[a] New entrants in 1960 defined as those born in a foreign country and residing abroad five years prior to the 1960 Census; new entrants in 1970 defined as those born in a foreign country and entering the United States in 1965–70.

Borjas notes the possibility of emigration selectivity, but presents no evidence of the importance of this form of selection. While we have no direct information as yet on the magnitude or direction of re-migration selectivity in terms of earnings, we can adduce evidence from the Census data used by Borjas on the extent to which entry cohorts of the U.S. foreign-born are altered due solely to emigration (or other selection processes) rather than to changes in the characteristics of individuals who stay (assimilation). We need only compare statistics on time-persistent characteristics of an entry cohort over time; any changes in these characteristics must be due to selection out (emigration), since, by definition, no cohort additions occur unless sample coverage rates increase. One characteristic that does not change with time (with few exceptions) for an individual, is not subject to recall error, and is not likely to be altered for reasons of ego, is gender.[10] Changes in the sex ratio of an age-entry cohort is thus likely to signal emigration, although it will surely underestimate it to the extent that emigration rates by sex are similar.

Table 4 reports the percentage change over two ten-year periods in the ratio of females to males for the foreign-born entering in 1955–60 and entering in 1965–70, by age group, with comparisons to the sex-ratio change for the native-born, taken from the 1960, 1970, and 1980 Public Use Tapes. As can be seen, the changes are considerably greater for the foreign-born entry cohorts in each age group (but one) compared to the native-born. While there is essentially no change in the sex ratio for the native-born in the (initial) age group 20–44 over either ten-year period, in both periods the ratios of females to males among the foreign-born cohorts rise considerably, by as much as 43 percent, except for the Western Hemisphere foreign-born who entered 1965–70. And, as is consistent with evidence on age effects on mobility, changes in the sex ratio tend to be greatest among the foreign-born in the youngest age group. Moreover, the changes differ significantly by Hemisphere and by decade, as would be expected given the differential changes in immigration law by Hemisphere in the 1960 to 1980 period. Table 4 thus demonstrates that emigration by the U.S. foreign-born is important and is selective by sex, by age, and by entry cohort. Interpretation of the effects of origin-country characteristics on the changes in the earnings of age-entry cohorts of the U.S. foreign-born thus is likely to require a richer model of self-selection, incorporating decisions to migrate to the United States and to remain there after immigration.

REFERENCES

Borjas, George J., "Self-Selection and the Earnings of Immigrants," *American Economic Review*, September 1987, *77*, 531–53.

Greenwood, Michael J. and McDowell, John M.,

---

[10] It is well known, for example, that retrospectively ascertained schooling attainment levels drift upward over the life cycle.

"The Factor Market Consequences of U.S. Immigration," *Journal of Economic Literature*, December 1986, *4*, 1738–72.

Jasso, Guillermina and Rosenzweig, Mark R., (1986a) "Family Reunification and the Immigration Multiplier: U.S. Immigration Law, Origin Country Conditions, and the Reproduction of Immigrants," *Demography*, August 1986, *23*, 291–311.

______ and ______, (1986b) "What's in a Name? Country-of-Origin Influences on the Earnings of Immigrants in the United States," *Research in Human Capital and Development*, Vol. 4, Greenwich, CN: JAI Press, 1986, 75–106.

______ and ______, *The New Chosen People: Immigrants in the United States*, New York: Russell Stage, forthcoming.

Self-Selection and the Earnings of Immigrants: Reply

By GEORGE J. BORJAS*

My 1987 paper addresses a central question in the economics of immigration: Who are the immigrants? It has been traditionally assumed that immigrants are positively selected—that is, immigrants are more skilled, on average, than the typical person in the source country. One contribution of my theoretical analysis was to show that this assumption is not necessarily consistent with wealth-maximizing behavior. In fact, under some conditions the immigrant flow may be negatively selected.[1]

My original study focused only on the selection bias caused by immigrants endogenously sorting themselves out of the source country's population.[2] Guillermina Jasso and Mark Rosenzweig point out that I ignored other types of selection biases in both my theoretical development and empirical framework. These additional biases are due to: (1) the selective emigration of the foreign-born in the United States; and (2) the fact that my empirical work was limited to immigrants originating in 41 source countries.

Regarding the first of these points, Jasso and Rosenzweig restate the well-known fact that emigration (or return migration) is an important characteristic of the immigrant experience. Robert Warren and Jennifer Marks Peck (1980), for instance, estimate that perhaps as many as 30 percent of immigrants leave the United States within a decade or two after their arrival. Jasso and Rosenzweig, however, do not examine systematically how the selective emigration of the foreign-born affects either the theory or the empirical analysis. In essence, their criticism simply asserts that selective emigration is likely to cause problems. My 1987 paper, however, already made note of that fact (p. 538). Moreover, the biases caused by emigration were an integral part of my earlier critique of the econometric methodology used by most studies of immigrant earnings determination (Borjas 1985, pp. 466–67). What is obviously needed is the generalization of both my theoretical and empirical frameworks to account for the emigration of the foreign-born. Such a generalization, however, is not provided by Jasso and Rosenzweig and does not yet exist.[3]

The second point made by Jasso and Rosenzweig deals with the empirical results presented in my paper. Because I limited my analysis to immigrants originating in 41 countries, the sample is not representative and serious selection biases may be at work. Obviously, it is always better to use a more representative sample in the empirical analysis. The data were limited to 41 source countries because I was concerned not only with immigrant wage levels, but also with immigrant wage growth as recorded by the 2/100 1970 and 5/100 1980 Public Use Samples of the U.S. Census. The relatively small sample sizes in the 1970 Census required that only countries with large immigrant flows as of 1970 be included in the study. Although these countries accounted for 90.4 percent of the immigrant flow between 1951 and 1980, they only accounted for 66 percent of the foreign-born who entered between 1975 and

---

*Department of Economics, University of California, Santa Barbara, CA 93106, and National Bureau of Economic Research. I am grateful to Bernt Bratsberg for providing me with the results reported in Table 1, and to the National Science Foundation, grant nos. SES-8604973 and SES-8809281 for research support.

[1] See Borjas (1990a) for an application of the conceptual framework set out in my 1987 paper to various aspects of the immigrant experience in the United States.

[2] I have recently generalized this model (Borjas, 1990b) to account for selection in both observables (such as education) as well as unobservable skill traits (which were the focus of my original paper).

[3] Borjas (1989) provides an empirical analysis of a longitudinal data set, the *National Science Foundation Survey of Natural and Social Scientists and Engineers*, that yields some insights into the biases caused by selective emigration of the foreign-born.

1980. This change in the nature of the sample represents the major shifts that occurred in the national origin mix of the immigrant flow during the past two or three decades.

Jasso and Rosenzweig are to be commended for attempting to generalize my empirical analysis by increasing the sample to 107 source countries (in Tables 1 and 2). In doing so, however, they misinterpret the implications of both my model and empirical results. In particular, given income maximization and lognormality in the earnings distributions, equation (6) in my 1987 paper shows that the average earnings of immigrants in the United States are given by:

$$(1)\qquad E(\ln w_1|\text{Immigrant}) = \mu_1 + \alpha[(\sigma_1/\sigma_0) - \rho]\lambda,$$

where α is a positive constant, and λ is the inverse of Mills' ratio ($\lambda \geq 0$). If earnings between the United States and the source country are positively and strongly correlated (i.e., ρ is near unity), positive selection is observed whenever the United States has more income inequality than the source country ($\sigma_1 > \sigma_0$), and negative selection is observed otherwise. It is useful to view the income inequality parameter as a proxy for the rewards to skills in a given labor market. In essence, the model says that prices matter: whether the immigrant flow is positively or negatively selected depends on the relative rewards to skills among countries.

In their discussion of equation (1), Jasso and Rosenzweig mistakenly define μ_1 as the mean earnings of natives in the United States. It is, instead, "the mean income that residents from the home country would earn in the United States *if all* home country citizens were to migrate to the United States" (Borjas, 1987, p. 532; emphasis in original). Generally, this mean income will differ among national origin groups. Even holding demographic characteristics constant, skills acquired in some source countries are more valuable in the U.S. labor market than skills acquired in others. It would not be unreasonable to expect that persons originating in economies resembling that of the United States (for instance, countries with high per capita incomes) have skills that are easily transferable and tend to be more successful in the U.S. labor market. Differentiating (1) with respect to mean income in the source country (μ_0) and with respect to mobility costs (π) yields:

$$(2a)\qquad \partial E(\ln w_1|\text{Immigrant})/\partial\mu_0 = \partial\mu_1/\partial\mu_0 + \alpha[(\sigma_1/\sigma_0) - \rho]\,\partial\lambda/\partial\mu_0,$$

$$(2b)\qquad \partial E(\ln w_1|\text{Immigrant})/\partial\pi = \alpha[(\sigma_1/\sigma_0) - \rho]\,\partial\lambda/\partial\pi.$$

Jasso and Rosenzweig assert that the derivatives in (2a) and (2b) should have the same sign regardless of whether there is positive or negative selection (because increases in μ_0 and π affect the probability of immigration and thus λ in the same way). As long as μ_0 proxies for the extent to which skills are transferable to the United States, the comparison of the signs of (2a) and (2b) says little about the type of selection that is being generated by the immigration market. As I concluded in my 1987 paper, the "earnings of immigrants in the United States may well be a positive function of μ_0" regardless of the type of selection characterizing the immigrant flow (Borjas, 1987, p. 536). Therefore, the Jasso-Rosenzweig "test" for positive (or negative) selection is based on a misinterpretation of the selection model.

In view of this fact, there is no need to address much of the discussion surrounding Jasso and Rosenzweig's Tables 2 and 3, except to note that they use the literacy rate as a measure of skill prices in the source country. Countries as diverse as Cuba, Japan, Sweden, and the United States have very high literacy rates, yet surely there is sizable variation in the returns to skills among these countries. Jasso and Rosenzweig, however, do not even state exactly why and how the literacy rate proxies for skill prices.

There are additional problems with the earnings functions reported in Table 3. Without explanation, 16 countries suddenly disappear from the data, so that their gener-

TABLE 1—DETERMINANTS OF STANDARDIZED WAGE OF NATIONAL ORIGIN GROUPS[a]

Variable	Coefficient
Intercept	−0.6218 (−3.72)
ln (Per Capita GNP in Source Country)	0.1122 (6.42)
Distance from U.S. (in 1000s of Miles)	0.0123 (1.41)
Dummy Indicating If Source Country Has a Communist Regime	−0.2235 (−3.55)
Dummy Indicating If Source Country's Income Distribution Has More Inequality than That of the United States	−0.0847 (−1.82)
R^2	0.563
Number of Observations	65

[a]The *t*-statistics are presented in parentheses. See Borjas (1987, Table 4) for a description of the variables.

alization of my empirical work is now based on 91 countries. More disturbing, the 50 additional countries increase the sample size in the individual based earnings function by only 585 observations. During the 1975–1980 period, Vietnam (one of the countries omitted in my work, but included in the Jasso-Rosenzweig analysis) accounts for over one-third of these additional observations. Hence the Jasso-Rosenzweig results are based on a regression where the typical additional country (other than Vietnam) is represented by fewer than 8 observations.

Moreover, although Jasso and Rosenzweig stress the fact that some of their *t*-statistics are more significant than mine, this observation too is inaccurate. As long as a common disturbance (or fixed effect) characterizes persons originating in a particular country, Brent Moulton (1986) shows that one must account for the intra-country correlation in order to estimate the standard errors of the country-specific variables. In effect, there are only 91 independent sources of information in the data (the number of countries), and not 2,534 (the number of individuals). Hence the OLS standard errors for the country-specific variables reported in Jasso and Rosenzweig's Table 3 are incorrect, and the *t*-statistics exaggerate the significance of the coefficients. Finally, although the literacy rate plays an important role in Jasso and Rosenzweig's discussion of Table 2, it is not even marginally significant in the earnings functions (even with its inflated *t*-statistic). Thus, much of their criticism about the admittedly inconclusive nature of my results is based on a regression that is, at best, questionable.

Despite these doubts about the specific contribution made by the Jasso-Rosenzweig analysis, their general point is quite valid and worth pursuing. Obviously, much more can be learned about the immigrant experience if additional source countries are included in the study. For instance, by restricting the analysis to the 1975–1979 cohort (as done by Jasso and Rosenzweig) there are 65 source countries which have over 30 working men enumerated in the 1980 5/100 Census and for which data on income inequality are available.[4] A log-wage regression was first estimated in the Census data (with 30,835 observations) which included a vector of demographic variables (including education, age, marital status, and disability status), and a vector of dummy variables indicating country of birth. The coefficients of these dummy variables were then regressed on the aggregate country-specific characteristics using generalized least squares. The resulting regression is presented in Table 1.

---

[4]See Borjas (1987) for a description of how the sample of working men is selected from the Census data and for a definition of the variables used in the regressions. In addition to the 41 countries used in my 1987 analysis, the regression in Table 1 includes Australia, Bangladesh, Belgium, Chile, China (as opposed to Taiwan), Costa Rica, El Salvador, Finland, Honduras, Hong Kong, Indonesia, Iraq, Kenya, Lebanon, Malaysia, Morocco, New Zealand, Pakistan, Peru, South Africa, Sri Lanka, Thailand, Turkey, and Uruguay.

An implication of the theory is that, holding migration costs and other characteristics of the income distribution constant (such as the means and the correlation coefficient), persons originating in countries that have more income inequality than the United States should earn less than persons originating in countries that have less income inequality than the United States. This prediction is confirmed by the results in Table 1. Holding source country income constant, immigrants originating in countries with substantial income dispersion earn about 8 percent less than immigrants originating in countries with little dispersion, and this difference is significant in a one-tail test at the 5 percent level.

Note that this result does not say that negative selection dominates the data. All it says is that immigrants originating in some countries are less favorably selected than immigrants originating in others. Despite the interpretation given to my results by Jasso and Rosenzweig, this is the fundamental point stressed in my paper, both conceptually and empirically.[5]

It is heartening to see the research strategy suggested by my work provoke discussion and further research. Immigration is (and will remain) an important policy problem, and the insights provided by the economic approach to immigration will surely play an important role in the ongoing debate.

REFERENCES

Borjas, George J., "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics*, October 1985, *3*, 463–89.

______, "Self-Selection and the Earnings of Immigrants," *American Economic Review*, September 1987, *77*, 531–53.

______, "Immigrant and Emigrant Earnings: A Longitudinal Study," *Economic Inquiry*, January 1989, *27*, 21–37.

______, (1990a) *Friends or Strangers: The Impact of Immigrants on the U.S. Economy*, New York: Basic Books, 1990.

______, (1990b) "Immigration and Self-Selection," in Richard B. Freeman and John M. Abowd, eds., *Immigration, Trade, and the Labor Market*, Chicago: University of Chicago Press, 1990.

Moulton, Brent R., "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, August 1986, *32*, 385–97.

Warren, Robert and Peck, Jennifer Marks, "Foreign-Born Emigration from the United States: 1960 to 1970," *Demography*, February 1980, *17*, 71–84.

---

[5] Borjas (1990b) tests for the presence of positive or negative selection in the data.

The Effects of Fiscal Policies When Incomes Are Uncertain: A Contradiction to Ricardian Equivalence: Comment

By WILLIAM A. BOMBERGER*

In the March 1988 issue of this *Review* Martin Feldstein presents a model in which Ricardian equivalence (hereafter RE) may fail to hold. For some parameter values, a present value-preserving transfer of income from future generations increases the consumption of the current generation. In addition to the potential failure of RE, the model has two other features: (a) members of the current generation must make some of their consumption decisions before their future income is known, and (b) for certain parameter values the failure of RE occurs even though some individuals leave altruistically motivated bequests.

It is well known that insufficient altruism or imperfect capital markets can cause a failure of RE. Likewise, the simultaneous failure of RE and occurrence of bequests is delivered by models which assume non-altruistic motives for bequests. Interest centers not on the fact that such results can occur but on identifying new and important factors which generate them. Accordingly, Feldstein emphasizes that aspect of the model which can (with appropriate parameter values) cause RE to fail: "the inability of individuals to predict accurately their subsequent earnings can cause a fiscal transfer from future generations to the current generation to raise current consumption" (p. 22).

However, Feldstein's claim, that the failure of RE in this model is caused by individuals' inability to predict future income, is misleading as a general statement and false when applied to the numerical examples Feldstein puts forward. Appropriate parameter values can generate a failure of RE in models with altruistic, rational consumers. The addition of unequal incomes or unequal degrees of altruism can produce bequests accompanied by the failure of RE. The assumed degree of altruism and distribution of income in Feldstein's numerical examples lead to a failure of RE by precisely this mechanism even when the inability to predict income is removed from his model.

Income uncertainty is neither necessary nor sufficient to produce a failure of RE in this model. Uncertainty is redundant for Feldstein's purposes in that all of the results he emphasizes can be generated by an appropriate choice of parameters in a perfect foresight version of the same model. These results are demonstrated below.

Feldstein's model is an overlapping generations model. Each member of the current generation lives two periods and has a utility function:

$$(1)\quad u = \ln(c_1) + \ln(c_2) + \alpha \ln(z_1 + z_2),$$

where c_1 and c_2 are the first- and second-period consumption of the current generation, $z_1 + z_2$ is the total lifetime income of the next generation, and α is a measure of intergenerational altruism. First- and second-period incomes for the current generation are y_1 and y_2, respectively. The real return on saving is zero.

Whether this setup delivers RE depends on income values and the degree of altruism embedded in the utility function. With identical households, RE will hold if and only if (all) members of the current generation leave bequests. If households are not identical, some may leave bequests while others do not. If so, bequests occur yet RE fails because part of any transfer from future generations produced by a debt-financed tax cut is consumed by households at corner solutions who leave no bequests. Barro (1976) mentions this case: "When households are not identical, the aggregate effect of government debt issue will depend on the fraction of households at a corner. As long as some households are in this situation, a shift (debt-financed tax cut) will have some... effect in this model. However, this effect would be "small" if the fraction of households at a corner is small." (p. 1104)

---

*Department of Economics, University of Florida, Gainesville, FL 32611.

Feldstein's model can generate an economy in which some individuals leave bequests yet RE fails in the aggregate by introducing inequality in $y_1, y_2, z_1 + z_2$, or α. Feldstein chooses to allow inequality in y_2. A fraction, p, of the first generation will receive a large second-period income, Y_2, and the remaining fraction, $1-p$, will receive 0. He assumes that those who receive Y_2 will leave bequests and those who receive 0 will not. He later selects numerical values consistent with this assumption. It is not surprising that the resulting model predicts that some bequests will occur and RE will fail.

The conventional source of this result is obscured because Feldstein also assumes that members of the current generation make first-period consumption decisions before they know to which second-period population they belong. They do not know who will receive Y_2 and who will receive 0. He then repeatedly asserts that the failure of RE which the model exhibits is caused by "income uncertainty."[1]

If uncertainty "causes" the failure of RE in a model, that failure should disappear if the model is purged of uncertainty while retaining its other elements. To indicate that this is not the case with this model I have reproduced Feldstein's numerical examples and extended them to include a setup where the "inability of individuals to predict accurately their subsequent earnings" is eliminated but all other aspects of the model, including the failure of RE, remain.

In the numerical examples Feldstein assumes $y_1 = z_1 = z_2 = 1$. A less altruistic ($\alpha = 1$) and a more altruistic ($\alpha = 3$) case are examined. For comparison, he examines (a) a "certainty" setup in which $y_2 = 1$ for everyone and, (b) an "uncertainty" setup in which $p = 0.5$ and $Y_2 = 2$. It is important to note that the "uncertainty" setup adds more than uncertainty to the model. First, individuals now receive unequal second-period incomes. Second, uncertainty is added by the assumption that second-period income is unknown when first-period consumption decisions are made. However, one could introduce inequality without uncertainty. This would be useful if one wishes to isolate the effect of uncertainty per se, which is the stated purpose of the exercise.

Accordingly, I add a third setup to this model. It is identical to Feldstein's "uncertainty" setup in that half of the current generation receive an income of 2 in the second period and the remaining half receive 0. However it is assumed that every member of the current generation knows in the first period what his (her) second-period income will be. This setup exhibits the same distribution of income as Feldstein's "uncertainty" setup while removing all aspects of the inability to predict future income which (according to Feldstein) is the source of the propensity to consume out of debt-financed tax cuts.

Table 1 presents some results for the three setups. Feldstein's "certainty" setup is labeled "identical" since all individuals receive identical incomes. I label my new deterministic setup "unequal" since there are two populations with unequal second-period incomes. Feldstein's second setup remains "uncertain." It adds uncertainty to the unequal setup by having all individuals make first-period decisions before knowing to which second-period population they belong.

---

[1] By assuming inequality only in second-period income and choosing the distribution he does, Feldstein creates a model in which households are identical *ex ante.* Choosing inequality in some other parameter or allowing households more information about their second-period income would change this while retaining the essential results of the model. It is not clear that Feldstein has deliberately sought to achieve this result nor is it clear that this is an advantageous or realistic feature of the model.

Feldstein has also chosen to make all uncertainty in the model individual or idiosyncratic rather than aggregate in nature. This is undoubtedly the most empirically important type of uncertainty. However, the addition of aggregate uncertainty could lead to different results. For a discussion of the importance of this distinction for RE see Andrew Abel (1988) and Louis Chan (1983).

Table 1—Numerical Examples

Setup	c_1	c_2	B	$\alpha = 1$ mpc_1	mpc_2	mpc	mpb
Identical							
(1)	1.00	1.00	0.00	0.50	0.50	1.00	0.00
Unequal							
($y_2 = 0$)	0.50	0.50	0.00	0.50	0.50	1.00	0.00
($y_2 = 2$)	1.50	1.50	0.00	0.50	0.50	1.00	0.00
(2) Average	1.00	1.00	0.00	0.50	0.50	1.00	0.00
Uncertain							
($y_2 = 0$)	0.63	0.37	0.00	0.58	0.42	1.00	0.00
($y_2 = 2$)	0.63	2.18	0.19	0.58	−0.29	0.29	0.71
(3) Average	0.63	1.28	0.09	0.58	0.07	0.65	0.35

Setup	c_1	c_2	B	$\alpha = 3$ mpc_1	mpc_2	mpc	mpb
Identical							
(4)	0.80	0.80	0.40	0.00	0.00	0.00	1.00
Unequal							
($y_2 = 0$)	0.50	0.50	0.00	0.50	0.50	1.00	0.00
($y_2 = 2$)	1.00	1.00	1.00	0.00	0.00	0.00	1.00
(5) Average	0.75	0.75	0.50	0.25	0.25	0.50	0.50
Uncertain							
($y_2 = 0$)	0.59	0.41	0.00	0.50	0.50	1.00	0.00
($y_2 = 2$)	0.59	1.10	1.31	0.50	−0.13	0.37	0.63
(3) Average	0.59	0.76	0.65	0.50	0.19	0.69	0.31

First- and second-period consumption and bequests are shown for each setup. For the unequal and uncertain setups, consumption and bequests are calculated for each income group and for the average of the two groups. Following Feldstein, I have calculated the change in current generation consumption and bequests produced by a 0.1 rise in first-period income linked to a 0.1 fall in the income of the next generation. The per unit changes in first- and second-period consumption are listed as "mpc_1" and "mpc_2," respectively. The per unit change in lifetime consumption, ($mpc_1 + mpc_2$), is "mpc." The per unit change in bequests is "mpb." Calculations are based on Feldstein's formulas (p. 19).

For the less altruistic case ($\alpha = 1$) with identical individuals (line 1), no bequests occur and RE is violated. Members of the current generation consume all of a debt-financed tax cut. Introducing inequality (line 2) does not change this. Even those for whom $y_2 = 2$ are not sufficiently altruistic to leave a bequest. Hence, the failure of RE is exhibited before the inability to predict future income is added. Adding uncertainty (line 3) causes not the failure of RE, but the occurrence of bequests. Due to additional first-period saving, those for whom $y_2 = 2$ are sufficiently endowed after their income is revealed to leave a bequest. The current generation as a whole will now consume about 65 percent of any debt-financed tax cut, rather than 100 percent. However, a debt-financed tax cut will cause the current generation to shift its consumption toward the first period and away from the second thus raising the first-period marginal propensity to consume.

For the more altruistic case ($\alpha = 3$) with identical individuals (line 4), everyone leaves a bequest and RE holds. The introduction of inequality in second-period incomes (line 5) is sufficient to cause the failure of RE even when second-period income is known. Those for whom $y_2 = 0$ leave no bequest and consume 100 percent of any debt-financed tax cut over their lifetime. The Ricardian behavior of those for whom $y_2 = 2$ brings the lifetime marginal propensity to consume down to 50 percent for the population as a whole. Adding uncertainty (line 6) changes magnitudes but does not change the pattern of bequests or the failure of RE.

Different parameter values for the "unequal" setup can generate models where: (a) all individuals leave bequests and RE holds, or (b) no individuals leave bequests and RE fails, or (c) some individuals leave bequests and RE fails. Adding uncertainty produces no additional cases. Uncertainty is neither necessary nor sufficient to produce (c), the "interesting" case (p. 18). Uncertainty is redundant in the sense that it only yields cases which could be generated by appropriate parameter values in a deterministic version of the same model.

Beside the question of which mechanism generates these results is the question of what implication these results have for the Ricardian controversy. Feldstein contrasts his model with models which generate similar results with non-altruistic bequest motives. The fact that this model generates a failure of RE without introducing such motives is taken as evidence of the "general inapplicability of RE" (p. 22).

It is hard to credit this inference. If one's goal is to establish that RE can fail in a theoretical model with altruistic individuals without market failures, that goal was achieved in Barro's model (among others) and is readily admitted by "Ricardians." The perspective of Ricardians is that the existence of such models is not controversial but is also not decisive for this question. One can present plausible models which support either position depending on parameter values which cannot be easily measured.

The major contribution of Barro's analysis is evidential. His model implies that the prevalence of bequests yields indirect evidence on the Ricardian controversy. If nearly all households leave bequests, near RE should hold. If few households leave bequests, substantial non-Ricardian results should occur. Rather than contradicting this central point, Feldstein's model has identical implications. In contrast, models with non-altruistic bequests overturn this implication by breaking the link between bequests and Ricardian behavior.

The introduction of uncertainty into this model for the purpose of demonstrating the potential failure of RE accompanied by bequests is unnecessary and unmotivated. The statement that uncertainty causes the failure of RE in this model is incorrect.

REFERENCES

Abel, Andrew B., "The Implications of Insurance for the Efficacy of Fiscal Policy," *Journal of Risk and Insurance*, June 1988, *55*, 339–78.

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November-December 1974, *82*, 1095–11.

Chan, Louis K. C., "Uncertainty and the Neutrality of Government Financing Policy," *Journal of Monetary Economics*, May 1983, *11*, 351–72.

Feldstein, Martin, "The Effects of Fiscal Policies When Incomes Are Uncertain: A Contradiction to Ricardian Equivalence," *American Economic Review*, March 1988, *78*, 14–23.

GAUSS THE CHOICE OF THE EXPERTS

“The speediest and most versatile of the mathematical programming languages.”

Barry Simon - PC Magazine

PC MAGAZINE EDITOR'S CHOICE MARCH 14, 1989

You get more work done when you use the right language.

Fast and Flexible

Now you can discover why scientists, engineers and statisticians agree that GAUSS is the language of choice when you want the most in speed and flexibility from your PC. The built-in matrix routines are written in assembly language and are faster than even the fastest C or Fortran compilers.

Multiple Regression

```
xxi = inv(x'x);
b = xxi*(x'y);
sse = (y-x*b)'(y-x*b)/(n-rows(b));
stderr = sqrt(diag(sse*xxi));
tvalue = b / stderr;
pvalue = cdftc(tvalue,rows(x));
```

It's not just execution speed that makes GAUSS the better choice. GAUSS is a matrix language, using familiar syntax for expressing mathematical relationships. The result is you will shorten your development time and write less code. GAUSS's programming environment has all the tools you will need to write large programs, including an on-line debugger and an on-line help system that can handle the functions you write.

Expandable

Create your own libraries of functions or call FORTRAN, C, or assembly language routines.

Publication Quality Graphics

The GAUSS system includes 2-D and 3-D publication quality graphics with an underlying resolution of 4096 x 3120.

Applications Modules

The optional SimGauss module simulates nonlinear differential equations, state-space and discrete systems. It's powerful, flexible and supports state/parameter vectors for increased speed and ease. Modules are also available for numerical optimization, solving simultaneous nonlinear equations, generating basic descriptive statistics, maximum likelihood, logit, probit, loglinear analysis, and others.

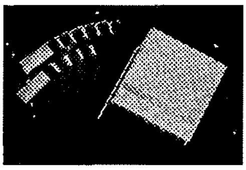
Economical

For just $495 you get the basic system including a library of mathematical functions and publication quality graphics.

Guaranteed

If you don't agree that GAUSS is the speediest and most versatile of the mathematical programming languages, simply return it for a full refund. No questions asked. Call (206) 631-6679 to order.

GAUSS–faster, easier, smarter... guaranteed.

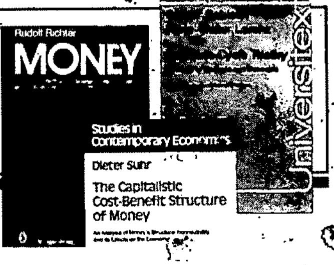


System includes 700 page manual and 11 program and application disks for IBM PC-XT-PS/2 & compatibles with math coprocessor. 386 protected mode version also available.

GAUSS™

BY APTECH SYSTEMS, INC.
26250 - 196th Place SE, Kent Washington 98042
Phone (206) 631-6679 FAX (206) 630-1220
CALL OR WRITE FOR FREE BROCHURE

Economic theory that's right on the money.


Rudolf Richter
MONEY
Studies in Contemporary Economics
Dieter Suhr
The Capitalistic Cost-Benefit Structure of Money
Universitext

☐ **The Collected Economic Articles of Harold Hotelling**
ed.by **A. Darnell,** University of Durham, U.K.
(1989, 178 pp., 10 illus., hardcover, ISBN 0-387-97011-8, $49.80)
Here are nearly all of the published articles in economics of one of the most successful twentieth-century pioneers of mathematical economics. Many economists today find inspiration in Hotelling's pathbreaking papers, and this collection should provide even further inspiration for research.

☐ **Business Cycle Theory:**
A Survey of Methods and Concepts
2nd Revised & Enlarged Ed.,
G. Gabisch & H.-W. Lorenz, both of University of Göttingen, FRG **(1989, 248 pp., 75 illus., softcover, ISBN 0-387-51059-1, $34.00)**
Presents a comprehensive survey of modern business cycle theory, encompassing shock-dependent and shock-independent business cycle models. This revised and enlarged edition includes a separate chapter devoted to new developments in dynamical economics.

☐ **Money:** ***Lectures on the Basis of General Equilibrium Theory and the Economics of Institutions***
R. Richter, University of Saarbrücken, FRG
(1989, 399 pp., 38 illus., hardcover, ISBN 0-387-50253-X, $90.00)
The concept of a currency order is at the center of this book, which develops monetary theory as a theory of currency orders — expanding the neoclassical theory of money along the lines of modern institutional economics.

☐ **The Capitalistic Cost-Benefit Structure of Money:** ***An Analysis of Money's Structural Nonneutrality and Its Effects on the Economy***
D. Suhr, University of Augsburg, FRG
(1989, 136 pp., softcover, ISBN 0-387-51138-5, $21.10)
Examines the dichotomy between real exchange and the monetary economy, offering the concept of a "neutral money" financial innovation.

You'll also want to consider these best-selling titles in economic theory:

Macroeconomics and New Macroeconomics
B. Felderer and **S.Homburg,** both of University of Cologne, FRG
(1987, 329 pp., 97 illus.,
☐ ***softcover:*** **ISBN 0-387-18004-4, $29.50;**
☐ ***hardcover:*** **ISBN 0-387-16961-X, $54.00)**

☐ **Microeconomic Theory**
Y. Otani and **M. El-Hodiri,** both of the University of Kansas, USA **(1987, 274 pp., 88 illus., softcover, ISBN 0-387-17994-1, $35.00).**

To order, call **1-800-SPRINGER** or 201-348-4033 in NJ; or send a copy of this ad with a check or AmEx, MC, or VISA number and expiration date, plus $2.50 for shipping, to: **Springer-Verlag New York, Inc.,** 175 Fifth Avenue, New York, NY 10010, Attn: F. Balzac-**Dept. S405.**

Name____________________________

Address____________________________

City/State/Zip____________________________

CENTRAL LIBRARY
13 FEB 1991

Springer-Verlag New York Berlin Heidelberg Vienna London Paris Tokyo Hong Kong

Springer for Economics. Depend on it.

THE AMERICAN ECONOMIC REVIEW

VOL. 80 NO. 2 MAY 1990

PAPERS AND PROCEEDINGS

OF THE

Hundred and Second Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Atlanta, Georgia

December 28–30, 1989

Program Arranged by Gerard Debreu

Papers and Proceedings Edited by Ronald L. Oaxaca and Wilma St. John

Copyright © AMERICAN ECONOMIC ASSOCIATION, 1990

CONTENTS

Editors' Introduction . . . *Ronald L. Oaxaca and Wilma St. John* vii
Foreword . . . *Gerard Debreu* viii

PAPERS

Richard T. Ely Lecture

Why Are We So Rich and They So Poor? . . . *David S. Landes* 1

Economic Education at the High School Level

Research on High School Economic Education . . . *William Becker, William Greene, and Sherwin Rosen* 14

The New Economics of Personnel: Rationale or Rationalization?

Firm Organization and the New Economic Approach to Personnel Behavior . . . *Debra J. Aron* 23
Who Bosses Whom? The Intensity of Supervision and the Discipline of Labor . . . *David M. Gordon* 28
Sticky Stories: Economic Explanations of Employment and Wage Rigidity . . . *Sanford M. Jacoby and Daniel J. B. Mitchell* 33

Aggregate Asset Pricing

Asset Prices under Habit Formation and Catching Up with the Joneses . . . *Andrew B. Abel* 38
Measuring the Persistence of Expected Returns . . . *John Y. Campbell* 43
Evaluating Empirical Tests of Asset Pricing Models: Alternative Interpretations . . . *Stephen G. Cecchetti and Nelson C. Mark* 48

Stock Market Volatility

Do Security Analysts Overreact? . . . *Werner F. M. De Bondt and Richard M. Thaler* 52
Market Volatility and Investor Behavior . . . *Robert J. Shiller* 58
Speculative Dynamics and the Role of Feedback Traders . . . *David M. Cutler, James M. Poterba, and Lawrence H. Summers* 63

The Economics and Politics of Budget Control

Budgetary Pressures in the EEC: A Fiscal Federalism Perspective . . . *David E. Wildasin* 69
U.S. Federal Budget Deficits and Gramm-Rudman-Hollings . . . *Edward M. Gramlich* 75
Public Debts and Fiscal Politics: How To Decide? . . . *Robert P. Inman* 81

The "New" Growth Theory

Trade, Innovation, and Growth . . . *Gene M. Grossman and Elhanan Helpman* 86
Why Doesn't Capital Flow from Rich to Poor Countries? . . . *Robert E. Lucas, Jr.* 92
Are Nonconvexities Important for Understanding Growth? . . . *Paul M. Romer* 97

Lessons for Development from the Experience in Asia

Lessons from Korean Economic Growth . . . *Susan M. Collins* 104
Asian Trade and Growth Lessons . . . *Anne O. Krueger* 108
External Sector in Development: China and India, 1950–89 . . . *T. N. Srinivasan* 113
Development Lessons from Asia: The Role of Government in South Korea and Taiwan . . . *Yung Chul Park* 118

Incentive Effects of Medical Malpractice
The Effects of Malpractice Litigation on Physicians' Fees and Incomes . . . . . . *Patricia M. Danzon, M. V. Pauly, and Raynard S. Kington* 122
Experience Rating: Does It Make Sense for Medical Malpractice Insurance? . . *Frank A. Sloan* 128

New Classical Macroeconomics in the Open Economy
International Transmission and Real Business Cycle Models . . . . . . *Alan C. Stockman* 134
Bargaining and International Policy Cooperation . . . . . . *Kenneth Rogoff* 139
The New Classical Macroeconomics and Stabilization Policy . . . . . . *Rüdiger Dornbusch* 143

The New Theory of the Firm
Equilibrium Short Horizons of Investors and Firms . . . . *Andrei Shleifer and Robert W. Vishny* 148
The Efficiency of Equity in Organizational Decision Processes . . . . . . *Paul Milgrom and John Roberts* 154
Asymmetric Information and the New Theory of the Firm: Financial Constraints and Risk Behavior . . . . . . *Bruce C. Greenwald and Joseph E. Stiglitz* 160

State and Local Government Finance
Inflation and the Real Growth of State and Local Government Expenditures . . . . . . *Roy Bahl and Jorge Martinez-Vazquez* 166
State Assistance to Local Governments: Changes During the 1980s . . . . . . *Helen F. Ladd* 171
Commercial Land Use Regulation and Local Government Finance . . . . . . *Eric A. Hanushek and John M. Quigley* 176

The Rationality of the Foreign Exchange Rate
Chartists, Fundamentalists, and Trading in the Foreign Exchange Market . . . . . . *Jeffrey A. Frankel and Kenneth A. Froot* 181
Volatility in the Foreign Exchange Market and Stock Markets: Is It Excessive? . . . . . . *Robert J. Hodrick* 186
Nonlinear, Nonparametric, Nonessential Exchange Rate Estimation . . . . . . *Richard A. Meese and Andrew K. Rose* 192

Interaction Between Environmental and Agricultural Policies: Opportunities for Coordination and Limitations for Evaluation
Interactions Between Agricultural and Environmental Policies: A Conceptual Framework . . . . . . *Richard E. Just and John M. Antle* 197
Coordinating Agricultural and Environmental Policies: Opportunities and Tradeoffs . . . . . . *S. R. Johnson, Robert Wolcott, and Satheesh V. Aradhyula* 203
Limitations in Evaluating Environmental and Agricultural Policy Coordination Benefits . . . . . . *James Hrubovcak, Michael LeBlanc, and John Miranowski* 208

Bargaining and Price Formation under Incomplete Information: Theories and Experiments
Explicit Communication and Bargaining Outcomes . . . . . . *Vincent P. Crawford* 213
Bad and Good News about the Sealed-Bid Mechanism: Some Experimental Results . . . . . . *Andrew Schotter* 220
The Transition from Bargaining to a Competitive Market . . . . . . *Steven R. Williams* 227

The National Research Council's Report on the Status of Black Americans, 1940–85
A Common Destiny: How Does It Compare to the Classic Studies of Black Life in America? . . . . . . *Thomas D. Boston* 232
Blacks, Hispanics, and White Ethnic Groups: Are Blacks Uniquely Disadvantaged? . . . . . . *Reynolds Farley* 237
The Central Role of the South in Accounting for the Economic Progress of Black Americans . . . . . . *James J. Heckman* 242
Racial Inequality in the Managerial Age: An Alternative to the NRC Report . . . . . . *William A. Darity, Jr.* 247

The Formation of Economic Values

Are Risk-Benefit Tradeoffs Possible in Siting Hazardous Facilities? . . . . *Howard Kunreuther and Douglas Easterling* 252
Sources of Inconsistency in Societal Response to Health Risks . . . . . . *W. Kip Viscusi* 257
Fixed Preferences and Changing Tastes . . . . . . *Edi Karni and David Schmeidler* 262

New Developments in Economic Theory

Information as an Economic Commodity . . . . . . *Beth Allen* 268
Evolution and Cooperation in Noisy Repeated Games . . . . . *Drew Fudenberg and Eric Maskin* 274
The Economic Implications of an Incomplete Asset Market . . . . . . *H. M. Polemarchakis* 280

Women's Labor Market Mobility: Evidence from the *NLS*

Child Care and Women's Return to Work After Childbirth . . . . . . *Jacob Alex Klerman and Arleen Leibowitz* 284
Intercohort Differences in Women's Labor Market Transitions . . . . . . *M. Anne Hill* 289
Gender Differences in Wages and Job Turnovers Among Continuously Employed Workers . . . . . . *Audrey Light and Manuelita Ureta* 293

Wage Trends and the Job Creation Debate

Earnings Inequality and Job Turnover . . . . . . *Jonathan S. Leonard and Louis Jacobson* 298
The Impact of Schooling and Industrial Restructuring on Recent Trends in Wage Equality in the United States . . . . . . *Barry Bluestone* 303
Schooling, Work Experience, and Wage Trends . . . . . . *Marvin H. Kosters* 308

Selectivity Bias: New Developments

Varieties of Selection Bias . . . . . . *James J. Heckman* 313
Nonparametric Bounds for Treatment Effects . . . . . . *Charles F. Manski* 319
Semiparametric Estimation of Selection Models: Some Empirical Results . . . . . . *Whitney K. Newey, James L. Powell, and James R. Walker* 324

Reviving the Federal Statistical System

Reviving the Federal Statistical System: The View from Academia . . . . . . *Jeffrey A. Miron and Christina D. Romer* 329
Reviving the Federal Statistical System: A View from Industry . . . . . . *Rosanne Cole* 333
Reviving the Federal Statistical System: International Aspects . . . . . . *Robert E. Lipsey* 337
Reviving the Federal Statistical System: A View from Within . . . . . . *Jack E. Triplett* 341

The Economic History of Technology

The Development of Industrial Research in U.S. Manufacturing . . . . . . *David C. Mowery* 345
Punctuated Equilibria and Technological Progress . . . . . . *Joel Mokyr* 350
The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox . . . . . . *Paul A. David* 355

Poverty and the Underclass

Effects of Family and Community Background on Economic Status . . . . . . *Mary Corcoran, Roger Gordon, Deborah Laren, and Gary Solon* 362
AFDC Participation Across Generations . . . . . . *Peter Gottschalk* 367
Welfare and Work in Chicago's Inner City . . . . . . *Marta Tienda* 372

Are We Saving Too Little?

International Differences in Saving . . . . . . *Barry P. Bosworth* 377
Public Debt as Private Liquidity . . . . . . *Michael Woodford* 382

Deregulated Airline Markets

The Dynamics of Airline Pricing and Competition . . . *Steven A. Morrison and Clifford Winston* 389
Airport Presence as Product Differentiation . . . . . . *Steven T. Berry* 394
Airline Mergers, Airport Dominance, and Market Power . . . . . . *Severin Borenstein* 400

Strikes, Bargaining, and Arbitration: New Developments

Can Strategic Bargaining Models Explain Collective Bargaining Data? . . . . *John Kennan and Robert Wilson* 405
Strikes and Bargaining: A Survey of the Recent Empirical Literature . . . . *David Card* 410
Negotiator Behavior and the Occurrence of Disputes . . . . . . *Orley Ashenfelter and Janet Currie* 416

The Industrial Organization of Markets with Differentiated Products

The Introduction of New Products . . . . *Debra J. Aron and Edward P. Lazear* 421
Pricing Behavior and Vertical Contracts in Retail Markets . . . . *Andrea Shepard* 427
Money-Back Warranties vs. Replacement Warranties: A Simple Comparison . . . . *Duncan P. Mann and Jennifer P. Wissink* 432

The AEA Commission on Graduate Education in Economics

Educating and Training New Economics Ph.D.s: How Good a Job Are We Doing? . . . . *W. Lee Hansen* 437
Discussion . . . . *Alan S. Blinder* 445
. . . . *Claudia Goldin* 445
. . . . *T. Paul Schultz* 447
. . . . *Robert M. Solow* 448

PROCEEDINGS

John Bates Clark Award 452

Minutes of the Annual Meeting 453

Minutes of the Executive Committee Meetings 455

Reports

Secretary . . . . *C. Elton Hinshaw* 461
Treasurer . . . . *C. Elton Hinshaw* 465
Finance Committee . . . . *C. Elton Hinshaw* 466
Editor, *American Economic Review* . . . . *Orley Ashenfelter* 467
Editor, *Journal of Economic Literature* . . . . *John Pencavel* 476
Editor, *Journal of Economic Perspectives* . . . . *Joseph Stiglitz* 479
Director, *Job Openings for Economists* . . . . *C. Elton Hinshaw* 481
Committee on Economic Education . . . . *John J. Siegfried* 483
Committee on the Status of Minority Groups in the Economics Profession . . . . *Ronald L. Oaxaca* 485
Committee on the Status of Women in the Economics Profession . . . . *Nancy M. Gordon* 486
Committee on U.S.–China Exchanges in Economics . . . . *Gregory C. Chow* 490
Ad Hoc Committee on the *Journal of Economic Perspectives* . . . . *William J. Baumol* 492
Representative to the National Bureau of Economic Research . . . . *David Kendrick* 496
Representative to the American Association for the Advancement of Science . . . . . *Adam Rose* 498
Policy and Advisory Board of the Economics Institute . . . . *Edwin S. Mills* 500

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the hundred and second annual meeting of the American Economic Association. The *Proceedings* record the business activities of the Association in 1989; the annual membership meetings; the March and December meetings of the Association's officers and committees. The *Papers* constitute the greater part of the volume. They comprise contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. We would like to take this opportunity to answer a number of commonly asked questions about the *Papers*.

Who chooses the authors? About a year in advance, the Association's President-elect, acting as program chairman, decides on the theme of the sessions and topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* was published annually in the Notes section of the *AER*, and now appears in the Fall issue of the *Journal of Economic Perspectives*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions will be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing 29 sessions, although a total of 96 sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies.

Are discussants' comments published? There has been no standard practice with regard to the publication of comments and discussions in the past. This year the President-elect decided to publish one set of comments for one panel session. For the other sessions, names and affiliations of commentators are printed at the start of each session, permitting readers especially interested in particular comments to write to the commentator for a copy of the discussion.

What standards must the papers meet? The guidelines under which papers are published in the *Papers and Proceedings* differ considerably from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions, and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to a formal refereeing process. However, a paper can be rejected if, after reading it, we conclude it is utterly without merit. This year we are pleased to report that no paper has been rejected on this ground. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

In order to produce this volume by May, very rigid deadlines must be met and there is no time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions. Every effort is made to notify an author prior to the deadline if the paper is too long, or does not satisfy other specifications.

This year, most authors cooperated very nicely. We thank them for making our lives easier.

RONALD L. OAXACA
WILMA ST. JOHN

Foreword

The task of arranging the program for the 1989 annual meeting of the American Economic Association and of publishing this issue of the *American Economic Review* was made possible by the willingness of many to contribute to its execution. To each one of them I owe personal thanks.

Rudiger W. Dornbusch, Stanley Fischer, Arthur S. Goldberger, Edward M. Gramlich, Daniel J. B. Mitchell, Alicia H. Munnell, Roy Radner, Richard L. Schmalensee, Eduardo S. Schwartz, and V. Kerry Smith, as members of the Program Committee, transformed a demanding duty into a rewarding experience. To them I am especially indebted.

GERARD DEBREU

RICHARD T. ELY LECTURE

Why Are We So Rich and They So Poor?

By DAVID S. LANDES*

At the risk of tipping my hand, I shall argue that most answers to the question posed by my title fall into one of two lines of explanation. One says that we are so rich and they so poor because we are so good and they so bad; that is, we are hardworking, knowledgable, educated, well-governed, efficacious, and productive, and they are the reverse. The other says that we are so rich and they so poor because we are so bad and they so good: we are greedy, ruthless, exploitative, aggressive, while they are weak, innocent, virtuous, abused, and vulnerable. It is not clear to me that one line of argument necessarily precludes the other, although most observers and commentators have a strong preference in the matter. What is clear is that, insofar as we may want to do something about the gap between rich and poor, each of these explanations implies a very different strategy.

I

In the beginning was Adam Smith, and he told us not to worry about economic growth: it would take care of itself. Left alone, people would sort things out, do what they did best, make appropriate choices to maximize return. The market would take care of the rest, rewarding reasons and quickness and knowledge and punishing the opposite. All of this, moreover, would work to the general advantage, augmenting the wealth of nations and leading them through a natural progression of stages from agriculture to industry to commerce. Long live the invisible hand!

To be sure, this sense of immense possibilities of improvement did not last. Malthus and Ricardo in particular developed theses of limits to growth that did much to earn for economics the name of dismal science. Malthus stressed the tendency in the long run for population to increase to and beyond the limits of subsistence, and linked this unhappy outcome to the inexorable operation of arithmetic. For Malthus, natural and man-made disasters—famine, disease, war—were the necessary winnows of a biosphere in disequilibrium. He was not a complete pessimist and recognized the small possibility that self-imposed restraint in reproduction might solve the problem, but given the force of human nature and the prevailing contraceptive technology (to say nothing of the absence of television and other compensating diversions), he was not very hopeful.

Ricardo took the stick from the other end: the limits to the extension of cultivation. As demand for food increased, he argued, ever-poorer land would be brought into cultivation, thereby raising the cost of food and wages, reducing profits, inflating rents, and crowding out other uses for capital. The motor of growth would simply seize up. The result would be the stationary state.

It would be rash to argue that Malthus and Ricardo were not an integral part of the classical paradigm. Yet their pessimistic lessons were in fact dismissed or, more precisely, were put away to be revived another day. In the heady days of nineteenth-century expansion, they seemed at best misguided. It is true that population was growing faster than before (although there were any number of misconceptions about what had passed before), but food was apparently no problem. On the contrary, the famines of yesteryear disappeared, for many reasons: new staple crops (the potato, maize); the application to the soil of outside nutrients;

---

*Department of Economics, Harvard University, Cambridge, MA 02138.

better rotations; virgin lands in and out of Europe; improved transport. And those regions where population pressed on subsistence were able to export their surplus eaters: the opening of frontier areas overseas seemed to provide an indefinite solution to the Malthusian dilemma.

As a result, the theme of limits to growth simply receded from the intellectual consciousness. Occasionally a lonely voice like that of Jevons was heard, warning of the exhaustion of the coal supply, but his anxieties were dismissed as parochial, especially after the technological innovations of the Second Industrial Revolution (liquid and gaseous fuels, electricity) transformed the economics of energy.

II

This "growth is natural" model (though no one would have called it that) remained for well over 100 years the dominant paradigm; so much so, that it became an invisible given of economic thought in general, and more or less disappeared as a subject of inquiry. Insofar as some nations had trouble following this path—doing what comes naturally—the explanation was as Smith himself understood it: man and politics had gotten in the way.[1] In particular, the intervention of the state, however well meant, worked to hobble initiatives, distort the market, and cripple the invisible hand.

The same sense of complacency prevailed in regard to distribution. Clearly some nations were richer than others. But that was all right because it was in the nature of things. Of the three factors of production—land, labor, and capital—it was the first that made the difference. Land (which included resources under the land and climate above) was unequally distributed. That was God's work. Those nations more richly endowed with resources were, other things equal (the saving proviso of economic thinking), bound to be richer. As for the other two factors, labor and capital, the assumption was that, in the long run, these were homogeneous and equal. People were rational maximizers or could like putty be shaped to the role; and money was money, subject to appropriate rates of exchange. Both factors were assumed to be mobile and/or elastic, ready to move to opportunity—labor by migration or population change, capital by transfer or saving. Even knowledge and know-how were there for the buying. Only land was different, and there, given this natural inequality, it was in the interest of each nation to make the best of what it had. Here the Ricardian analysis of comparative advantage reinforced the Smithian model and the contentment that went with it.

And the discontent. It should not be hard to understand that those countries dissatisfied with their place in the economic order were not prepared to accept this inequality as God-given, that is, to take the message of the economic pundits as gospel. The wealth of nations was also the power and place of nations. Very early on, then, indeed from the eighteenth century, those countries such as France that could reasonably presume to rival Britain understood that Britain's lead in technology was a source of potential dominance as well as wealth and sought to follow suit. To this end, they sent emissaries to look and learn; and where these were barred by business secrecy, they sent spies. They also did their best to lure British workers and mechanics, and to import machines for use but, above all, for copying. The British in turn, who clearly did not believe in the assurances of their best and brightest about the virtues and rewards of laissez-faire (at least not where exports were concerned) issued laws and decrees to bar the sale of machines and the emigration of artisans. To no avail: in the long run, like love, the market laughs at locksmiths.

Sad to relate for sound economic policy, the would-be follower countries were not content to confine themselves to importing and stealing knowledge and hardware. Unwilling as they were to allow cheaper British goods to swamp their own industries, they threw up trade barriers, and these proved more enforcible and effective than British export embargoes. (To limit fraud, for example, the French made use of outright prohibitions: that way, any piece of foreign cloth had to be contraband and was ipso facto subject to seizure.) The British, in turn, who had their own long history of protection and interference with trade, deplored this perturbation of the natural order and regretted the unwillingness of other countries to join them in a spirit of shared freedom and international cooperation. John Bowring, for example, Parliamentary emissary extraordinary, lamented in his report of 1840 the rising German duties on manufactures—Germany, he pointed out, so clearly unsuited by nature to industry and destined by its resources to serve as a granary for Albion's Workshop of the World.

---

[1] See, on this point, Smith's analysis of China's "stationary" state: the country "had probably long ago acquired that full complement of riches which is consistent with the nature of its laws and institutions. But this complement may be much inferior to what, with other laws and institutions, the nature of its soil, climate, and situation might admit of" (p. 95).

Even this was not enough. These follower countries, determined to help themselves, began taking steps to hasten the adoption of the new techniques. They recognized perfectly that backwardness had its servitudes and exigencies. They understood, for example, that they needed more capital than the British to pay for newer vintages of machinery and would have to invent institutions to mobilize these resources; and that they had to take deliberate measures to train technicians and engineers to deal with technologies they had not grown up with at the bench and in the shop. No one told them, as a later generation of economic historians would, that steam power didn't matter if water power was available; that if they didn't have iron, they could always use wood; that they were already doing as well as or better (growing as fast as or faster) than the British; indeed, that it was only chance that had given Britain its head start and that the Industrial Revolution might just as easily have taken place in France; and besides, even if Britain were ahead, what difference did it make, since a mere 15 percent gap in wages or income (if not more) was just not enough to fuss about; besides, *steak pommes frites* tastes better than fish and chips.[2]

What's more, even if some clever cliometrician had given them the good tidings, using stochastic models and statistical constructs and such clever bits of folk wisdom as the multiple ways to skin a cat, these politicians and functionaries would have told them to get out of their book-lined studies and look around, and reminded them that political power has its own reasons and its own urgency. Instead, then, of consoling themselves with a most estimable array of cliometricians (how could they?), the decision makers of the follower countries read and heeded Alexander Hamilton, J.-A. Chaptal, Friedrich List, Christian Rother, and Gustav Mevissen. And so these countries caught up with Britain, to Britain's slow but enduring disbelief.

In all fairness, their ability to catch up was the result as much of preparedness as of deliberate preparation. The cultural and intellectual distance that separated them from Britain was relatively small; in some important areas of science, chemistry for example, they were well ahead; and their own technical and commercial experience was far from trivial. They had their own history of invention and manufacture, their own specialties, their own areas and niches of market dominance. And what they did not have, they could get from one another. The history of British, then French and Belgian and Swiss, then German export of capital and enterprise in the nineteenth century would have rejoiced the heart of any practitioner of marginalist economics.

By the end of the nineteenth century, then, the Industrial Revolution that had begun in Britain had diffused throughout Europe and to European offshoots overseas. Not to Latin America, whose monied elites were long content to trade primary products for overseas manufactures (in those days, such things as wheat, meat, coffee, and copper; today much the same, plus cocaine); nor to European colonies or even free countries in Asia. Africa, especially sub-Saharan Africa, lay beyond the pale of awareness. And if one had asked a European economist about this, he would have described it once again as the natural order of things. The international division of labor had been modified by the diffusion of the new technologies. Britain was no longer the Workshop of the World, which had expanded to include Europe and the United States. But specialization remained, and no European would have seen it as anything but rational and logical, inscribed in geography and, for many in that era, in the racial endowment.[3]

---

[2] For a dismayingly credulous survey of what passes as the latest revisionist interpretation of the Industrial Revolution, see Eric Jones (1988, ch. i): "A Know-all's Guide to the Industrial Revolution." On the substitutability of wood for iron and steel, in refutation of the "iron and steel propaganda of the Industrial Revolution," see F. T. Evans (1982); on the "dethroning" of the role of steam (Jones, p. 17), see Dolores Greenberg (1982); on the relative rates of British and French growth, see Richard Roehl (1976; 1983), and Patrick O'Brien and Caglar Keyder (1978); on the role of chance in the British primacy, see N. F. R. Crafts (1985); on wage, income, and culinary gaps, see O'Brien-Keyder and Donald McCloskey (1988).

III

It should be said that the Marxian mode of analysis, which presented itself as a dissent, did not reject or seek to alter this paradigm. Marx accepted the naturalness of growth and the positive link between technical advance and wealth. His primary concern was with distribution and the class relations implicit in or derived from the ownership and use of the means of production. In this regard, he never asked questions about the reasons for or determinants of technology or the mode of production—and this for the very simple reason that he did not think there was anything to explain.

The pursuit of productivity was a perfectly natural effort to enhance relative surplus, thereby increasing the rate of exploitation and promoting the accumulation of capital. Besides, for Marx, technological innovation was typically the work of science, and as such, the gratuitous fruit of what he called social labor. The capitalist was a taker. The first to take (to innovate) generally went broke, and the legacy was culled by the jackals of enterprise. "It is, therefore, generally the most worthless and miserable sort of money-capitalists who draw the greatest profit out of all new developments of the universal labour of the human spirit and their social application through combined labour."[4] So much for innovation.

What Marxists (as opposed to Marx and to mainstream economics) did come to concern themselves with was the distribution of wealth among nations. (One can see the analogy to his preoccupation with the question of distribution within the economy.) Specifically, they raised the issue of the economic consequences of imperialism and likened the workers of colonial countries (what we would now call Third World countries) to an external proletariat exploited by the capitalists of imperial states. This analogy, which offered polemical advantages in the attack on the old order, had the additional merit for some Marxists of explaining in politically congenial terms such evidence of improvement as they were willing to concede in the standard of living of workers under capitalism. The argument ran that the capitalists were buying off their own workers and thereby buying peace at home by squeezing these outsiders—in effect, by draining their surplus. At the same time, the imperial powers were said to be deliberately holding back these subject countries, blocking their industrial development for the benefit of capitalists at home.[5]

---

[3] The one element that should have given pause was the rapid growth and transformation of Meiji Japan, and this in spite of severe constraints on commercial autonomy imposed by the European powers on the pre-Meiji shogunate around the middle of the century. (The Japanese could not set protective tariffs to shelter their new industries.) It was their resounding defeat of Russia in 1905 that gave the measure (sounded the alarm) of the changing economic and political order in the Far East. One reaction was to adopt the Japanese as spiritual Westerners, "like us." People will do almost anything to save a paradigm.

[4] Karl Marx, *Capital* (III, ch. v, sect. 5, p. 103). See M. M. Bober (1968, pp. 282–83).

[5] This is a condensed, homogenized version of a diverse literature. Marxist critiques of imperialism range from simple land-drain arguments—it is a system for growing the cash crops and emptying the mines with cheap, forced or semiforced labor of a kind one could not employ at home—to conspiracy theories about capitalist unwillingness to help colonies develop industrially in competition with the mother country. For a

(Note that this thesis was sharply at variance with Marx's own favorable view of the economic consequences of empire: he castigated its abuses but considered it "objectively" progressive. Nothing else, he thought, could shake the torpor of centuries, otherwise known as the Asiatic mode of production. The thesis also clashed with the stereotype of capitalists who supposedly have no loyalty to anything but money. Why should they prefer dear labor to cheap?)

Intellectually, however, the Marxist anti-imperialist thesis remained in the master's tradition by accepting the classical paradigm; that is, it saw growth as natural and interpreted colonial poverty and backwardness as products of political exploitation and the wrong kind of interference. It assumed, therefore, that once the burden of captivity removed, the liberated colonies would be free to pursue their own destiny and develop economically as had their rich oppressors.

IV

When non-Marxist, "mainstream" economists belatedly began looking at the question of growth and development in what we now call the Third World, they were no more inclined than the Marxists to jettison the classical paradigm. The Western experience was proof of what could be done, even by countries that seemed destined to serve as sources of primary products. Canada and Australia, even Argentina and Brazil: there was no reason why a nonindustrial country could not eventually create a balanced, diversified modern economy. What it needed was good government and good markets, and resources would flow to the areas of highest return. If some of these went to industry, say, food processing, why that was just fine, especially if such movement reflected true marginal rates of return and not distorted rewards. Staples theory (vent for surplus) was invented to explain this process, and it seemed to work well with a variety of economies in time and space, ranging from Canada (furs, timber, grain, minerals), to the United States (tobacco, cotton, grain), to Sweden (timber, copper, iron ore), and perhaps eventually to Argentina (hides, grain, frozen meat), and Brazil (gold, sugar, hardwoods, coffee), and even to medieval England (wool).

The trouble was that once the development bug bit, the poor countries of the twentieth century had no patience for the slow, selective, and contingent success of staples growth. On the contrary, they saw it as a trap. In this they were really no different from the follower countries of Europe in the eighteenth and nineteenth centuries. Like them, they were in a hurry, and if anything in a greater hurry, because they were poorer and, thanks to the demonstration effect, hungrier. (If you ask any of the follower countries today whether they are prepared to wait 100 years to catch up, they will express outrage. Yet that is how long it took Japan.) The primary producers of the twentieth century found that most staples were easily substitutable and subject to fierce competition in world markets; hence that staples income was uncertain and beyond their control. They also found, as nineteenth-century exporters had, that private revenue from staples exports enriched disproportionately a small fraction of the society, who more often than not were self-indulgent consumers of luxury imports, who preferred rents to the risks of market competition, and who therefore avoided engagement in a broader pattern of development.

At the same time, these would-be developers were not prepared to eschew industrialization, that is, to accept the apparent dictates of comparative advantage, because industry, especially heavy industry (above all, coal, steel, and machines), spelled power and Marxist theory told them that there could be no modernization without what

---

guide to some of this diversity, much of which takes the form of rediscovering old news, see Keith Griffin and John Gurley (1985). Note, of course, that this criticism of imperialism was system-specific; that is, it applied only to capitalist or bourgeois government, not to socialist regimes. (Most Marxists would describe socialist imperialism as a contradiction in terms.) In this regard, the Marxists, as collectivists, were far closer to the Tories than to classical economic liberals. Compare S. M. Lipset (1988, p. 30).

Marx called Modern Industry.[6] (In all this, they had the example of such earlier developers as the United States and Japan, which may have built their earliest gains on light industry but then shifted resources into such branches as metallurgy.)

The result was Third World development economics, which bore a strong resemblance to its intellectual predecessors of the nineteenth century (Hamilton, List et al.), but modified, first by Marxian notions of the primacy, indeed the indispensability, of industry; of the superior if not sole legitimacy of government or collective ownership of the means of production, including peasant land; and of the importance of state planning and intervention; and second, by post-Marxian concepts and grievances of international exploitation and the penalties of inequality.

V

It should be noted that Third World (like First World) awareness of backwardness (or of poverty, for that matter) came relatively late. It was a twentieth-century phenomenon and was closely related, first, to the growing movement for freedom in such colonies and dominions as India (Indian ideological leadership in this regard was testimony to the already extensive transfer of British ideas and ideals along with technology), and then to the achievement of independence in the aftermath of World War II.

In logic and essence this new consciousness was, as we have seen, hopeful and even confident, for the lifting of the foreign thrall, with all the exploitation and impoverishment that that presumably entailed, could not but inaugurate a happier age of initiative and progress. Building on the assumption that late is better (that it pays to be late), economists and statesmen were quick to predict high rates of growth that would enable the new follower countries to catch up in short time with their richer predecessors. Such is the power of compound interest.[7]

These sanguine predictions have been disappointed; and that raises the question why. Let me start with the premise: the assumption that growth is natural, fairly waiting to happen.

The experience of the twentieth century, and especially of the period since World War II, is that development is neither natural nor easy. We have seen that, on the basis of the conventional wisdom, whether classical, neoclassical, or Marxist, it ought to be. Certainly the incentive is there: the gap between what is and what can be is enormous. And the opportunity. Once (here fill in the appropriate condition) the burden of colonialism is lifted, the government sets growth as the objective, the plans are drawn up, and the requisite resources are mobilized, growth and development should follow as the night the day. Only it has not.

To be sure, it is not hard to find grounds for encouragement. There has been staple growth of the classical variety in some of the oil-producing countries, specifically those with large deposits and small populations. More important, there are some instances of incipient but highly promising industrial growth. The best examples here are the East Asian little dragons (Hong Kong, Singapore, Taiwan, and Korea), but coming up behind are lesser beasties such as Thailand and perhaps Malaysia, and one should not overlook those awakening giants such as Brazil, India, or Turkey that have now passed from the Third to the Second-and-a-half World (or is it the First-and-a-half World?). Any repeated visitor to these countries can cite gains in the material standard of life, in the buildup of social overhead capital, in the provision of health care and welfare, in life expectancy.

But these achievements are not unmixed or unambiguous. Things may be better than they were, but they leave much to be desired, especially in those countries that are still in between. Just look at the morbidity data; or better yet, the people and the way they live.

---

[6] Compare in this regard the Soviet plans of the 1920s and 1930s and the Chinese plans of the 1950s and 1960s. (See Fureng Dong, 1988, pp. 235–36).

[7] One can read calculations (in effect quantified predictions) by economists and politicians of democratic republic X or ex-colony Y, newly launched on its populist path, showing it duly catching up with capitalist countries A, B, and C by the year 19.... (See, for example, Surendra Patel, 1962.)

Even the showcase examples give cause for anxiety. Take the new rich. Staples booms in wasting assets such as oil are of their nature ephemeral. Even while they last, they bring in windfall wealth that may or may not be a boon, depending on what one does with it. Economists have recently been wont to call this kind of problem the Dutch disease, after the disruptions to the Netherlands economy caused by the discovery of natural gas in the North Sea. But do not worry about the Dutch; they have always known what to do with money. Economic historians, I think, would prefer to speak of the Spanish disease, after the unhappy consequences of American bullion for the Spanish economy of the sixteenth and seventeenth centuries.

As for the instances of industrial growth, which are often cited as harbingers, they vary widely in character and scope. The evidence shows that industrialization has been a heavy and not always supportable strain on resources; that it has been a source of serious disamenities (pollution in particular, nowhere worse than in the Second and Third Worlds) that do not show in the income data; that it has produced societies that are deeply split between old standards and new, old ways and new (in some respect, things seem to get worse before they get better).

Much of this development, moreover, is externally rather than internally generated. It is based on a marriage of foreign capital and know-how (multinationals, joint ventures, offshore production) with cheap domestic labor. To be sure, capital has always been drawn to cheap labor, and much of early European industrial development can be accounted for in similar fashion. (This was particularly true of the diffusion of new technologies.) But such imports must be precarious and superficial until they are absorbed by the host society and converted into indigenous initiatives. Otherwise they remain a kind of industrial plantation.[8]

Finally, we should note that the growth experience is still relatively short, and this in countries where the fragility of political structures is an invitation to reversal. We should be cautious, then, in our conclusions.

And these are the winners. Against them one can cite a large number of examples of countries that are not keeping up, that are losing ground relatively and often absolutely, and this in spite of large inflows of capital and other resources. Some of this is obviously the consequence of man-made affliction: civil war; intertribal massacre; political mismanagement. But much of it is deeper and more enduring than the hazards of events. Sub-Saharan Africa and parts of Asia live day in and out with hunger, disease, and natural disaster so constant as to be endemic. Africa in particular, the region of most rapid population increase, is a heartbreaker, and there is as much reason to believe that its people are losing the battle for improvement as that they are winning it.[9]

VI

So the picture is mixed; to the point where it is now a commonplace to note that the Third World is a heterogeneous congeries of nations, rich and poor. Some of the distinction is based on the localized distribution of windfall staples wealth; some of it on real differences in the ability to absorb new technologies and grow. Whatever the source of the distinctions, one has the sense of a conceptual unity in course of dissolution. It is coming apart. Some countries are being "promoted," as it were, into the ranks of the advanced, industrial nations. Others are trying very hard and are still in midstream. Still others are for the moment getting nowhere.

---

[8] For this connection, see Jan Fagerberg:

> Both this logic and the subsequent test point strongly in the direction of a close relation between economic growth and growth of national [indigenous] technological activities, a fact that is normally forgotten. Thus, to catch up with the developed countries, the results obtained here suggest that semi-industrialized countries cannot rely only on a combination of technology import and investments, but have to increase their national technological activities as well. [1988, p. 451]

[9] The region of most rapid population increase: the implications for growth and development are not necessarily negative. Population is like raw capital: its value depends on what one does with it. Used well, it can be an asset, an incentive to investment, the human material of industrialization. Used badly it can be a crushing burden, a seedbed of emergencies and hasty improvisation, a source of political instability. It would be unwise to prejudge these matters, but the historical experience does tell a very different story for Europe (and Japan, for that matter), where rates of growth were never so high (a third perhaps), where the birth rate in the most advanced countries varied in response to the demand for labor (i.e., jobs chased people); and where the areas of population surplus were able easily and legally to export people elsewhere. Europe effected its population transition during its industrial transformation. The Third World today is growing more populous in advance of development, and most rapidly in the regions of slowest development.

In effect, we have the glass half-full, half-empty. Some would argue from success that all it takes is to get things right: wise policies; true prices. Others would argue from failure that getting things right is never an accident, and that some (many) may be condemned to persistent lateness and hence relative if not absolute failure.

What are the implications? Is it merely a question of optimists versus pessimists?

The question needs to be reformulated. We are talking about late development, of semi-industrial and preindustrial nations that want to catch up with a process of growth that began over 200 years ago. Well, does it make any difference to be a late developer? Or, to put it differently, does it pay to be late? (One proviso by way of circumscribing the question: I shall not concern myself with the losses in potential income incurred by lateness, i.e., by the difference between what has been and what might have been. Instead, I shall focus on the character and extent of the gains that follow on the inception of the development process.)

The conventional wisdom has always been that lateness is an advantage; that the gap between what is and what can be is a tremendous opportunity; that the follower country can profit from the experience and knowledge of its predecessors and avoid their mistakes; and that by mobilizing resources and allocating them energetically to the right uses, it will in fact grow faster than its forerunners. This was the argument made by Alexander Gerschenkron in his seminal articles of 1951–52 on "Economic Backwardness in Historical Perspective." Gerschenkron based his analysis on the European experience, on the comparison and contrast among Britain, Germany, and Russia in particular, and offered a "spurt" model of late growth.[10] He noted, to be sure, that such spurts, when driven from above (i.e., by the state), could impose a heavy burden on the population, to the point of exhaustion; hence the Russian pattern of alternating surges and collapses. But given good judgment and management, there was no reason why a follower country could not catch and even surpass its predecessors.

(And yet the European experience already gave grounds for pause. Russian growth was spotty, with large areas of persistent backwardness; and the experience of a socialist economy has only hardened this unbalanced pattern. Eastern Europe in general, in spite of advanced regions and local centers, was and remains a mixed story. The same has been true for Mediterranean Europe, where the more successful industrializers such as Italy suffered and indeed continues to suffer from dualistic contrasts in performance—and this in spite of extensive and deliberate efforts to promote the development of backward regions. We have in effect a pattern of incomplete modernization that has since become a commonplace.)

That was Europe in the nineteenth century. Some of the experience since then would seem to support the Gerschenkron thesis. Thus the high growth rates of such countries as Taiwan and Korea (7 and 8 percent per capita over a period of decades) show that it can still pay to be late. These are economies that have passed very rapidly through an import-substitution phase to export-led growth, much of it in the newest, most technology-intensive branches. Who would have thought it possible?

---

[10] Although himself politically conservative, Gerschenkron's spurt model has close similarities to the Soviet economic plans of the 1920s and 1930s and their ideological premises. Compare Stalin's emphasis on high-speed growth, on "overtaking and outstripping the advanced capitalist countries." "...the tempo must not be reduced. On the contrary, we must increase it...." The same for the 20th National Congress of the CPSU in 1961: "...efforts must be made to augment the industrial output by 2.5 times in the next decade and surpass the level of the industrial development of the United States." The same for Communist China, which aimed at "catching up with and surpassing Britain in the output of steel and other major industrial products" in 15 years or less. Similar parallels between the Soviet emphasis on heavy industry and the Gerschenkron model. (See the article by Dong, pp. 235–36.)

On the other hand, the moderate success of others and failure of still more have led some to argue that lateness is now a growing handicap. The reasons for such a judgment are not far to seek:

1) The size of the gap. It is now a gulf and keeps widening. By the older paradigm, that only means bigger potential gains to change. On the other hand, the threshold costs are higher. Capital is not the biggest problem. Knowledge and know-how are more esoteric, even opaque, hence harder to come by. Two possibilities present themselves: (a) hire people; (b) train one's own people. The former is expensive, and the best usually have better things to do. So one makes do with less than best (LTB), which may be less than enough. The second is also expensive, not so much for the cost of training as for the permanent loss of talent. How ya gonna keep 'em down on the farm after they seen Paree, London, Cambridge, Berkeley, or what have you? Again, the best are the ones with the least incentive to return; again, one can settle for LTB, which may or may not be enough.

2) Staples are not what they used to be. The same technology that has produced this inequality of nations works to limit the market power of primary products by making them more substitutable. Take sugar, a commodity of unusual potency in economic history. There was a time, in the eighteenth century, when this luxury-become-necessity could provide the basis of French commercial prosperity and of the industrial growth of the western half of the country: fleets, ports, *fabriques*, all hanging on the cane crop of one island, Saint-Domingue. By the beginning of the nineteenth century, however, that was over: France had been cut off by war from overseas supplies and had learned to make sugar from beets; while other centers of cane cultivation had developed to replace what was now Haiti, lost to sugar and to France as a result of the world's first successful slave revolt.

One could tell similar stories about rubber, food crops, even rare minerals.

3) Lateness makes for bad politics. It creates uncomfortable pressures, which conduce to poor answers. This has always been true, but at one time these pressures were the exclusive concern of governing elites: the ordinary Frenchman of the late eighteenth or early nineteenth century was not aware of and could not have cared less about industrial and technological changes across the Channel. In the twentieth century, however, awareness has been enhanced by the demonstration effect, itself much reinforced by new media of communication; and political urgency has been aggravated by ideological conviction and commitment. Governments are expected to deliver, to their own members to begin with, to the populace thereafter.

Hence great haste, with much waste. Lateness is the parent of bad government. Economists have been quick to point to the adverse effects of bad government on development (indeed, some would call it the primary cause of development failure), but have said little about the sources of bad government itself, which they see as properly the matter of other disciplines. Yet bad government—or for that matter, any kind of government, good, bad, or indifferent—is not unrelated to economics.

(To be sure, some of it is: most of the Third World countries are new, inexperienced excolonial nations, inheritors of irrational, accidental boundaries and of immemorial tribal or ethnic hostilities; also of authoritarian imperial structures that often flout older tribal patterns of discussion and consensus and facilitate obstinate error and abuse; heirs also of capital installations that they were not always in a position to maintain [there is nothing so demoralizing and demeaning as decay and abandonment]. So historical accident has played a role.) (Compare George Ayittey, 1989.)

Government is clearly part of a larger social system that includes economic structures and relations. (Marxists, indeed, would go farther and say that it is the creature of class relations and interests.) Good government is not there for the wanting, or even for the knowing. It is not an act of will or fiat. It will not come about because someone appoints good counselors, even good economists—who may well be our students and who, like us, may or may not agree. (And even if they did, most politicians would say that business and the economy are too important to be left to the economists.) It takes time to create an effective, functional bureaucracy; also to establish a commitment to a larger national identity and purpose. European countries took centuries to do this; new nations have tried to establish the whole panoply of institutions in a matter of years or decades. It is no accident that the success stories of East Asia are of relatively homogeneous societies with a strong sense of historical and cultural identity.

For new nations, moreover, the process has been immensely complicated by the grievances stored up over years of subordination and humiliation; by egalitarian ideologies that deprecate private success while justifying public privilege; by the impatience to set things right and catch up...quickly, NOW; by the choice of the fast and meretricious over the slow and steady; by the ubiquity of the state, which distorts the reward pattern and makes it easier to get rich by politics than by industry, by connections than by performance; and by the interplay or private, rent-seeking interests that are only too quick to exploit these possibilities.

4) Misdiagnosis and mistreatment. There's nothing that succeeds like success, and conversely. Lateness ideologized is like a malady that invites, even seeks out, bad therapy.

When Gerschenkron wrote about this problem, he offered the undisprovable thesis that nations would leap the gap between backwardness and development when they were ready. Today, by one definition, every nation is ready; and when things do not work out, they do not console themselves with the thought that they have been untimely. Rather they look for villains, whom they characteristically find outside themselves. (They can hardly be expected to blame themselves; besides, it would be morally wrong to blame the victim.)[11]

I need not go into the detail of these alleged sources of failure. They are familiar to all of us: colonialism or neocolonialism, unequal trade, underdevelopment (a noun derived from a newly invented transitive verb, to underdevelop), peripherality, dependency.[12] There is some truth in all of these, and with will and good will, there is much that can be done to eliminate or mitigate their effects. On the other hand, they are more the symptoms than the explanation of development failure. There are few of these alleged sources of backwardness, for example, that do not apply to Korea or Taiwan, both formerly Japanese colonies, both deliberately pastoralized by their rulers. And many of them apply to the British colonies in North America, even to the early American republic, and to Meiji Japan. All of them reflect circumstances of inequality that yield to sovereignty and to performance: make a better, cheaper radio, TV, watch, etc., and the world will be happy to do business with you on equal terms.

---

[11] This pattern of explanation goes back to the earliest industrialization efforts of what we now call Third World countries. I'm thinking, for example, of what was surely the first push of its kind, that of Mohammed Ali in Egypt in the period from about 1820 to 1848. Mohammed Ali was determined to bring Egypt into the modern world by effecting an industrial revolution. To this end he hired foreign specialists, imported equipment that he had copied as well as installed, established schools to train a cadre of technicians and professionals—the whole project conceived in the largest terms on the basis of European models. But the society was not ready for this leap. Mohammed Ali was unwilling to pay the necessary premiums to attract voluntary wage labor. (The Europeans, facing similar reticence, found the answer in using women and children, the people who could not say no; but a Muslim society was severely limited in this regard.) He began by recruiting slaves and, when too many of these died, moved on to corvée labor. Forced labor is unhappy labor, and soon sabotage became a serious drain on productive capacity; the favorite instrument of complaint was arson. Machines lay idle for want of maintenance and lack of parts. The whole system was breaking down by the time Britain, in 1839, succeeded in defeating Mohammed Ali's military ambitions and imposed a trade regime that limited Egypt to a modest revenue tariff. No protection. Subsequent generations of Egyptian historians have wanted to place the blame for the Mohammed Ali experiment on this constraint.

[12] The last three of these doctrines have come to us from Latin America, which, because it has been independent for a century and a half, has a special problem with the common recourse to neocolonialism as an excuse for failure. The difficulty is compounded by nature's bounty: these are lands generously endowed by nature that were able as a result to achieve considerable staples growth. At the turn of the century, Argentina, for example, was widely seen as a nation of unlimited possibilities, destined soon to take its place among the richest in the world. A half-century later, however, it was clear that none of these countries had done much to convert these earnings into balanced growth, including industry, so that the years of "follow-up" and catching-up were still ahead.

What's more, even if this bill of indictment were true, it would not pay to dwell on it. It leads to self-pity, myopia, and counterproductive policies. At the extreme, it would suggest complete delinking and economic isolation. Also, there is nothing so self-defeating as the transfer of responsibility and blame to others, if only because there are limits to altruism. After an initial surge of guilt, generosity wanes; it is a wasting asset. Indeed, the greater the benefit to others of unequal arrangements, the less likely they are to surrender them. The market, like God, best helps those who help themselves.

4) Cultural factors. Values are an especially thorny problem for would-be developers, partly because, insofar as they are an impediment to growth, they are strongest in "traditional" societies; and partly because they tend to be reinforced by economic failure. To be sure, economists do not like these. They lie outside the purview of the discipline, and they always seem to get in the way. (Historians, on the other hand, to say nothing of sociologists, have often cited them as explanations for exceptional economic performance in earlier periods [compare Max Weber and *The Protestant Ethic*]; or for Japanese achievements today.) They are often rejected as implicitly immutable, almost congenital (hence racist), although there is nothing to that effect in the argument. Or they are rejected for just the reverse, as epiphenomena that will yield easily to interest (in both senses of the word) and reason.

The truth, as so often, lies somewhere in between. Values and attitudes do change, but slowly, and their force and influence vary with circumstances. Many religious values operate, for instance, to impede the mobility and openness conducive to efficient allocation of resources and rational economic behavior. Worse yet, insofar as economic development entails changes in social structures and relations, vested cultural values, like vested material interests (they are in effect interests), can become a potent force for resistance, to the point of overturning governments and reversing the course of development. These setbacks, moreover, can be self-reinforcing, for the same cultural values that helped bring them about are also a precious source of consolation. The worse things get, the more the clock turns back, the more some people cling to what they know and feel and need.

VII

In the meantime, the struggle to pass from preindustrial to industrial, from "backward" to "advanced," goes on. By that I do not simply mean growth in income per head. That would be too easy. "Intensive growth," as it is sometimes called, can come about because nature has been kind, because new crops are more productive than old, because new land (including resources) becomes available, because relative prices change, because of outside developments and a free ride. But sustained growth is not possible without technological progress and gains in productivity. And that, history tells us, requires sooner or later the creation or assimilation of new kinds of knowledge and organization, which in turn depends on transformations within the society. External, enclave development will not do.

Such transformations require not only the absorption and adoption of new ways, but also, for many societies, the creation and acceptance of a new ethic of personal behavior. New ways demand and make new people. Time consciousness must become time discipline; the organization and character of work, the very relations of person to person, are transformed. These changes do not come easy. Historically they were often achieved by building on the more docile members of the society, the ones who could not say no, that is, on women and children, and that way creating a new labor force over a period of generations. This is still true. They have been most readily effected in those societies, like the Japanese, which had already developed appropriate time and work values before the coming of modern industry.[13] Selection, then, is not a matter of chance or need or desire.

So the transition to modernity is necessarily a case-by-case process. Many try but few are chosen. Insofar as the transition is adventitious, superficial, or forced, moreover, it proves to be discouragingly fragile, at least in the early stages. (This is especially true of windfall staples growth: witness the experience of Cote-d'Ivoire.) Small wonder that development is full of mistakes and disappointments, or that what seems like a breakthrough often slows or aborts.

Which raises another question: if promotion is slow and if the gap between rich and poor keeps growing, especially for the hindmost, and with it the obstacles, do we not have to change the paradigm? Are we not in fact facing, not a general and inevitable life process of maturation, however protracted, but rather a partially stochastic process of selection with diminishing chances of success? As time passes, those most qualified make it; but those who do not make it lose ground and become less and less qualified. They are the hard cases. They may, like the poor of the Bible, always be with us.

Where does that leave us? We are not going to give up, for that would go against one of the deepest values of our civilization: the Faustian urge, indeed the need, to shape our destiny, and everyone else's. For their good, of course. As for those who are not moved by altruism, they should recognize that it is not in their interest to allow extreme differences to subsist. We want to, we also have to improve the condition of the laggards. In the long run, of course, it is cheaper to do so by helping them to help themselves, but in the meantime, we must do for them. (As the old proverb puts it, it is better to teach a man to fish than to give him fish; but the quickest way to put fish on the table is to put it there.) They may not catch up, they may continue to lag, but they can become much better off. So long as we do not succeed in this, as now for example, we shall find that they export violence and people (i.e., their substance) instead of goods and services.

There is a popular children's book called *The Phantom Toll Booth* (by Norton Juster) that tells the story of a young boy engaged in a heroic quest. Before he sets out, he is told that there is one thing he ought to know, but it were best he be told after his return. So the lad sets forth, accompanied by friends, and after terrible dangers returns with mission accomplished. At which point he is told what he should have known all along, that the task was impossible.

We have the task. It is ours as well as theirs. It is impossible. But we must act as if it were not, for the reward is in the trying. They must too, for more is better than less. It may no longer pay to be late, but better late than never.

REFERENCES

Ayittey, George B. N., "The Political Economy of Reform in Africa," *Journal of Economic Growth*, Spring 1989, *3*, 4–17.

Bober, M. M., *Karl Marx's Interpretation of History*, 1st ed., 1927; Cambridge: Harvard University, 2nd ed., rev., 1968.

Crafts, N. F. R., *British Economic Growth during the Industrial Revolution*, Oxford: Clarendon, 1985.

Dong, Fureng, "Development Theory and Socialist Developing Economies," in Gustav Ranis and T. Paul Schultz, eds., *The State of Development Economics: Progress and Perspectives*. Oxford: Basil Blackwell, 1988.

Evans, F. T., "Wood since the Industrial Revolution: A Strategic Retreat?," *History of Technology*, 1982, *7*, 37–56.

Fagerberg, Jan, "Why Growth Rates Differ," in Giovanni Dosi et al., eds., *Technical Change and Economic Theory*, London and New York: Pinter, 1988.

---

[13] On the precocity of Japanese sensitivity to the efficient use of time, see Thomas Smith (1986). As early as the eighteenth century, instruction manuals for peasants, presumably literate, laid stress on the value of anticipation, preparation, and filling the temporal envelope.

Greenberg, Dolores, "Reassessing the Power Patterns of the Industrial Revolution: An Anglo-American Comparison," *American Historical Review*, December 1982, *87*, 1237–61.

Griffin, Keith and Gurley, John, "Radical Analyses of Imperialism, the Third World, and the Transition to Socialism: A Survey Article," *Journal of Economic Literature*, September 1985, *23*, 1089–1143.

Jones, Eric L., *Growth Recurring: Economic Change in World History*, Oxford: Clarendon, 1988.

Lipset, S. M., "Neoconservatism: Myth and Reality," *Society*, July-August 1988, 29–37.

McCloskey, Donald, "The Storied Character of Economics." *Tijdschrift voor Geschiedenis*, 1988, *101*, 643–54.

Marx, Karl, *Capital: A Critique of Political Economy*, Vol. III, F. Engels, ed., Moscow: Foreign Languages Publishing, 1894/1962.

O'Brien, Patrick K., and Keyder, Caglar, *Economic Growth in Britain and France 1780–1914: Two Paths to the Twentieth Century*. London: Allen & Unwin, 1978.

Patel, Surendra J., "Rates of Industrial Growth in the Last Century, 1860–1958," *Economic Development and Cultural Change*, April 1962, *9*, 316–30.

Roehl, Richard, "French Industrialization: A Reconsideration," *Explorations in Economic History*, July 1976, *13*, 233–81.

______, "British and European Industrialization: Pathfinder Pursued?," *Review* (Fernand Braudel Center, SUNY-Binghamton), 1983, *6*, 455–73.

Smith, Adam, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 1st ed., 1776, New York: Modern Library, 1937.

Smith, Thomas, "Peasant Time and Factory Time in Japan," *Past & Present*, May 1986, *111*, 165–97.

Wrigley, E. A., *People, Cities and Wealth: The Transformation of Traditional Society*. Oxford: Basil Blackwell, 1987.

ECONOMIC EDUCATION AT THE HIGH SCHOOL LEVEL[†]

Research on High School Economic Education

By WILLIAM BECKER, WILLIAM GREENE, AND SHERWIN ROSEN*

Twenty years ago, George Stigler argued that "economics belongs in everyone's education once we have learned how to teach it," but he concluded that economics is not easy to learn or teach: "Economics is not yet ready to be made a part of the basic curriculum of all educated men" (1970, p. 80). Stigler's conclusion was not accepted by the education establishment and the teaching of economics in high schools expanded greatly in the 1970s and 1980s. Ignoring the most obvious question of "why," here we address what has been learned about the teaching and learning of economics in high schools. Attention is given first to the ways in which learning has been measured, and then to research findings. Thoughts on the use of alternative measures of "value-added" and problems of sample selection are considered later. With few exceptions, our review is limited to the literature written for economists. We are extremely brief in the consideration of all topics, with a more detailed discussion to appear in *The Journal of Economic Education* (Summer 1990).

I. Measurement of Knowledge

Business, political, and labor leaders share the view that the public does not understand economics, although they may not agree on what constitutes an understanding of economics. Recently, John Soper and William Walstad advanced the notion that high school students do not know economics. With the support of the Joint Council on Economic Education (JCEE), Soper and Walstad (1987) developed the *Test of Economic Literacy* (*TEL*), which is a nationally normed 46-item multiple choice test. Walstad and Soper (1988) report that their selected sample of 2,483 students averaged only 20.34 questions correct, in a midschool-year administration of the *TEL*. At a news conference at the 1988 American Economics Association (AEA) meetings, Walstad called these results "shocking," since 11.5 questions correct could be expected from just guessing on each question with four alternatives.

But are these results shocking? The standard deviation was 7.45 answers correct, which implies a standard error of 0.15 questions and little likelihood of obtaining these results if students were just guessing. Before concluding that students know economics, however, recognize that the *TEL* was constructed so that each question tends to identify students who know from those who do not know economics. Reliability is built into the test by using questions that differentiate the same students in the same way; students with higher scores on the test should get any question correct with higher probability than those with lower scores. This rule for inclusion makes questions dependent, and the probability of successful guessing on each question is not fixed at 0.25. Whether any percent correct is higher or lower than what is expected depends on the test construction.

[†] *Discussants*: William Baumol, Princeton University and New York University; William Walstad, University of Nebraska.

*Indiana University, Bloomington, IN 47405; New York University, New York, NY 10006; and University of Chicago, Chicago, IL 60637. We benefited from the reviews of Dawson (1977), Wentworth, Hansen and Hawke (1977), Siegfried and Fels (1979), Siegfried and Walstad (1989), Becker and Walstad (1987), Watts (1990), Highsmith and Baumol (1989), and the annotated bibliography provided by the National Center for Research in Economic Education, University of Nebraska. Comments were provided by S. Becker, G. Dawson, R. Fels, K. Goldberg, P. Grimes, W. L. Hansen, R. Highsmith, P. Kennedy, T. Kniesner, J. Lopus, N. Peterson, P. Saunders, J. Siegfried, W. Walstad, and M. Watts.

In addition to sponsoring the development of the *TEL*, the JCEE directed a survey of high schools in the 1986–87 school year. This data set, known as the National Assessment of Economic Education (NAEE), involves information on 121 classrooms and 3,266 high school students (William Baumol and Highsmith, 1988). The NAEE had only one administration of the *TEL* as a year-end test; the average was 20.7, which is slightly lower than the 22.14 year-end average reported by Walstad and Soper (1988) for their sample of 2,483 students. The NAEE data also provide extensive qualitative information on students, teachers, and schools. Forty-three scholars were introduced to this NAEE data base, the Walstad-Soper *TEL* data base, and econometric modeling in economic education research (Becker and Walstad, 1987) at Princeton University summer workshops in 1987 and 1988; much of the recent research reported here is the outcome of those workshops.

II. Research Design and Findings

One might be tempted to subtract the midyear average *TEL* score from either of the two end-of-year averages reported above as a measure of what is learned in a half year of high school. This difference ignores numerous covariates that influence economic learning and knowledge of economics. To assess economic learning, researchers have used a "change score model" in which the post-test minus pretest score is the dependent variable and explanatory variables reflect the effects of individual characteristics (race, sex), human capital (IQ, prior grades, attitudes), utilization rates (study and class time), and technology (alternative teaching methods or equipment). Others use an "attainment model" in which the postmeasure is regressed on the pretest (when available) and the explanatory variables to assess economic knowledge. With these two models, researchers have learned the following about efforts to raise student learning and knowledge of economics.

A. *Special Projects*

Projects administered by Junior Achievement, chambers of commerce, and trade associations are aimed at increasing student understanding of certain aspects of an economic system, or of specific institutions within that system. Few studies of the outcomes from these projects have been undertaken and none have appeared in the academic, refereed journals typically read by economists. In contrast, there is a substantial body of published research on the Developmental Economic Education Programs (DEEP).

Michael Watts (1990) discusses the projects that have been conducted as part of DEEP and Watts (1985) provides an analysis of a statewide program. The DEEP projects share a common goal of raising high school student understanding of basic economics, although no uniform curriculum is dictated. Participating districts are expected to cooperate with a JCEE-affiliated state council or a collegiate center for economic education in two basic functions: conducting in-service programs for teachers (typically consisting of economics and education courses) and building economics into the curriculum through formal economics courses and other courses as well.

The many studies of DEEP show that high school students do learn economics in various types of courses and that teacher in-service programs are beneficial to student learning. Sherrie Rhine (1989), however, reports the curious results that DEEP is effective in states that mandate course work in economics, but it has a negative influence on student performance in states without a mandate. Rhine's results may be due to sample selection and variable definition problems, however. Whether a school district is or is not in DEEP may be of little importance compared to how successful it is in getting students and teachers into the study of economics (Becker and Walstad, 1990).

Norris Peterson's (1989) NAEE study supports this conclusion. It appears that it is the two basic functional parts of DEEP (teacher training and curriculum change) that are important and not the imprimatur.

B. *Student Ability*

Research confirms that the higher a student's aptitude or intelligence, the greater the learning in economics. The pretest score is typically the single most important variable in explaining attainment. It also influences change scores, although the bias caused by correlation between the error and a pretest regressor in these specifications has been ignored by researchers.

Student attitudes toward economics are thought to affect the learning of economics. Early work at the elementary grade levels suggests that the direction of causality is from learning of economics to positive student attitudes towards economics and not the reverse (as reviewed by Walstad, in Becker-Walstad, 1987). This direction of causality is supported by Kurt Beron's (1990) NAEE results that student performance influences student interest more than wanting to take more courses affects performance. These results may be sensitive to model specification, since they are based on attitude attainment and not changes in attitude. They suggest, however, that if students are to learn economics, we should first figure out how to teach it; they will then develop an appreciation for the subject.

C. *Teacher Ability*

There is a positive relationship between teacher knowledge of economics and learning of economics by students in their classroom (Andrew Weaver et al., 1987). Similarly, teacher attitudes toward economics may influence student attitudes toward economics (Howard Schober, 1984; James Marlin, forthcoming). Stephen Buckles et al. (1984) report that high school teachers do not want in-service programs on economics content as much as in-service programs that present teaching materials. Teachers may be able to learn economics if it is taught through materials that teachers could use in their high school classroom (Daniel Thorton and George Vredeveld, 1977).

William Bosshardt and Watts (1990) report that teacher training in economics and the quality of the students are important factors in student learning of economics. In both fixed and random effects models, they found student learning to be associated with instructors. In the fixed-effects model, for example, 32 of the 0–1 covariates included to represent the 94 instructors were significant. Jane Lopus (1990) reports that student knowledge of economics in the NAEE data is significantly related to expenditures for teachers with advanced degrees, smaller class sizes, and economic education consultants for teachers; student knowledge is not related to teacher experience. Gerald Lynch (1990) also notes that the first few courses in economics that a teacher takes may contribute little to the teacher's high school students' learning unless those students are being taught in an economics course (as opposed to a history or social studies course). Inherent in assessing the influence of the teacher, however, is the problem of separating the teacher from the environment.

Paul Grimes and Charles Register (1990) report that in the NAEE data, high school students in unionized schools score higher on the *TEL* than students in nonunion districts. Grimes and Register suggest that while their results are consistent with those of R. Eberts and J. Stone (1987), where a positive effect of teachers' unions on learning in elementary schools was found, a unique and compelling rationale is lacking. Clearly there is more to learn about the role of teachers in economic education.

D. *Course Work*

Fifteen states mandate a separate one-semester course in economics for high school graduation; one requires a course for college-bound students, four more states require that a course be offered as an elective (Highsmith, 1989). It simply is assumed that any course work in "economics" increases economic literacy, regardless of the content of the courses or the training of the teachers. In the case of mandated courses, however, Marlin's work calls this assumption into question.

Students can learn economics in high school economics courses, as well as in social studies, history, and business courses, where the teacher is trained in economics. John Chizmar et al. (1985) report that learning in economics even improves students' skills in other social studies. Different sets of prepackaged classroom materials have been shown to work in these varied settings (Chizmar and Ronald Halinski, 1983; David Martin and David Bender, 1985). George Dawson (1982), for example, reports particularly favorable results when economics is "infused" into an American history course. Walstad and Watts (1985), however, questioned how many students are exposed to any significant economics instruction through infusion. In consumer courses, learning of economics might even be negative according to Walstad and Soper (1988). Peterson reports that in the NAEE data, better students are avoiding economics courses. He is dubious of conclusions from studies of learning in various types of courses when such selection bias is not modeled.

E. *Technology*

Computers are the latest in a long string of technologies that have been touted as potential replacements for classroom teachers and for producing a major change in education as we know it today. In the 1950s and 1960s, television, programmed learning, and flash cards were considered revolutionary. In the 1970s, television and the show and tell approach of Sesame Street were pushed as the wave of the future. Games and simulations also caught on. Now we hear that interactive computing and interactive television are the future of education.

To date there is little research that shows any one of these existing technologies superior to another, or indeed that anything is superior to the classroom teacher. W. Lee Hansen et al. (1970) hypothesized that finding of no significant average differences ignored distributional effects. No researcher has followed up on the notion of matching teaching techniques to individual high school student attributes and benefits from learning economics. This may be because court decisions and social and political pressures within the United States make homogeneous ability groupings, tracking, and other forms of curricular selectivity difficult in high schools. These issues, however, need to be addressed by researchers.

Physical capital may be a poor substitute for human capital in precollege teaching. Rosen (1987) argues that in the economics of education industry, the student-teacher ratio becomes the most important constraint. Teaching and learning require a two-way communication that becomes increasingly unidirectional as the scale of the media is enlarged. Thirty years ago, the AEA-CEE and JCEE cooperated in the development of a national television course, *The American Economy*. Although this series had over a million viewers daily and was shown to be effective as a teaching tool (Phillip Saunders, 1964; and G. L. Bach and Saunders, 1965), it did not give rise to additional programming and did not revolutionize the teaching of economics. If such mass media teaching is desired by students and society, after thirty years why hasn't it made the small classroom obsolete? Why do parents continue to lobby for smaller classes and not more TV courses?

The conclusion that no superior technology has emerged does not imply that none could be used. Alternative technologies can complement the classroom teacher and allow for more individualized instruction. Just as *The American Economy* TV series was shown to be instructive, more recent TV series such as the JCEE's *Give and Take*, *Economics U$A*, and *Understanding Taxes* can be used to teach secondary students (Chizmar et al.). The question is, at what cost? Computers can complement classroom instruction (Darrell Lewis et al., 1985), but is it worth the cost to have each student sitting at a terminal? Furthermore, microcomputers may have made obsolete all previous evaluations of the role of mainframe computers in education. There is no research yet that shows any great breakthrough in the educational process. Taking a cue from George Stigler, Rosen

(1987) claims that there has not been a major innovation in education since the invention of the printing press, which made inexpensive book publication possible.

F. *Sex and Race*

Conventional wisdom holds that males have a relative advantage in numerical skills while females excel in verbal skills. This has given rise to an extensive literature on gender differences in learning economics.

John Siegfried's (1979) summary of earlier work suggests that high school males know more economics than females, as measured by multiple choice tests. Watts (1987) shows that males have higher levels of economic understanding as early as grade five. Males and females tend to learn the same amount during high school and college, so the gap that materialized before high school never closes during the schooling experience. The recent work of Keith Lumsden and Alex Scott (1987) suggests that this gap may even increase in the college economics course.

Although studies show a statistically significant gap in male and female learning of high school economics, the difference is typically small in magnitude—on the order of a couple of multiple choice questions, when background variables are controlled. An exception to this finding is Julia Heath (1989), who reports that male students outperformed women by 10 *TEL* points in economics courses after controlling for the fact that women do not elect to take economics classes. The work of Peterson suggests, however, that Heath's results may be sensitive to the peculiar way in which those who did not take economics had their *TEL* scores censored.

Regardless of the magnitude of the difference, the work of Marianne Ferber et al. (1983) and Lumsden-Scott suggests that observed gender differences may be the product of the multiple choice instrument. Women tend to do better than men on essay exams. Lumsden and Scott state: "One cannot avoid the uneasy feeling that the now generally accepted 'truth' of female inferiority [in the learning of economics] is based on ad hoc rationalizations" (p. 366).

As with gender, the effect of race on student learning of economics is not great. In Stephen Buckles and Vera Freeman (1984), for example, whites show a slight but insignificant advantage in learning economics in seven of the ten precollege grades considered. In their national study, Walstad and Soper (1989) found that blacks score little more than one question lower than whites on the *TEL*, with statistical significance depending on the model specification.

G. *Age and Time Usage*

Older students know more economics and are able to learn more abstract concepts than younger students. Buckles-Freeman and Watts (1987 and earlier articles) demonstrated, however, that students younger than 18 years of age can learn economics.

While there are no studies that explicitly measure time on task at the precollege level, the study by Stephen Jackstadt and Christian Gootaert (1980) suggests that working 20 hours or more per week substantially reduced knowledge and learning of economics. Supposedly students who spend so much time working cannot devote sufficient time to schoolwork. Females, however, appear to be less adversely affected by part-time employment, for some unknown reason. Jane Lillydahl (1990) found that modest levels of employment during high school may not interfere with student academic performance, but that students who work in excess of 20 hours per week are absent more often from school, spend less time on homework, and have lower academic performance scores. These findings stand in contrast to some of the general education research that shows high school grade point averages unaffected by part-time employment, (Gary Green and Sue Jaquess, 1987). The influence of time usage on student learning is not yet settled.

H. *Lasting Effects*

High school courses in economics do raise the economic understanding of students, at least in pre-post testing. Still, the lasting effect of high school courses in economics has been debated for years (see Siegfried and Rendigs Fels, 1979). Students who have had an economics course in high school enter their first course in college economics with a greater understanding, yet upon completion of that course they know little if any more than those who never took a high school course. Anthony Myatt and Charles Waddell (1990) report similar results, except the decay in learning does not show up until the student reaches the intermediate courses. The reason for this decay in learning is not clear, but one can speculate that students with high school economics spend less time on college economics to achieve a target grade and spend more time in other activities.

The new Advanced Placement Program in Economics, developed by the College Board and the Educational Testing Service and being promoted by the JCEE, enables high school students to earn college credit for economics learned in high school (Buckles and Morton). Whether or not this Advanced Placement Program will yield positive effects on student learning of economics as they move through college remains to be seen. Past experience with high school courses is not encouraging, however.

III. Problems with Value-Added

Most research in economic education has been based on test scores. The differences in these test scores are used as the "value-added" by the treatment. The market for new high school graduates does not place a value on student learning; it values the final level of accomplishment. Employers buy terminal high school aptitudes and skills, not a change in scores. What the student knew four years ago in the first semester of the freshman year is irrelevant to the employer, except insofar as it affects the rate of learning. Knowing the difference between test scores is of little help to a student or society without knowing the value of this difference.

Changes in test scores may have administrative value to a principal or classroom teacher, but that may have little relationship to the economic concept of value. Just as water has a high "value in use" but a low "value in exchange," some basic skills, such as an ability to reconcile a check book, may have high value in use but low value in exchange. Other skills may have a high value at one point in time and little value at another; for example, an ability to manipulate a slide rule has fallen in value since the availability of the cheap hand calculator. Although some skills may be viewed as essential for education, their market value is determined by demand and supply. The normative beliefs of a principal, school board, chamber of commerce, state legislature, or school faculty about the importance of intellectual skills is elusive without reference to what employers are paying for the bundle of skills embodied in the high school graduate, and what skills they desire from the graduate. (The satisfaction derived from learning and its change score measurement is ignored here for brevity, but is modeled in Becker, 1982.)

Hansen et al. called attention to the problem of valuing individual student learning and its implications for curriculum reform but few have followed their lead. Researchers can answer questions about the effect of courses in economics on the decision to go to college or on the probability of majoring in economics. There is no limit to the ways in which knowledge of economics might influence the decision-making processes and be related to observable outcomes.

IV. Problems of Data Loss

In addition to the problem of valuing changes in scores, there is a practical problem of measuring change: students "drop out" between the pre- and the post-test, and the process that determines who drops out is likely related to the process that determines scores. This data loss may bias estimators in attainment and change score models (Becker and Walstad, 1990).

Missing observations on key variables can also devastate a large data set. In the Walstad-Soper (1988) *TEL* data, 18 percent of the 3,031 matched pre-post scores were discarded because of missing data on other variables; in Walstad-Soper (1989), the matched prepost sample had grown to 3,427 but the usable sample had shrunk to 1,630 records. In the NAEE data, questionnaire data are missing on one-third of the teachers and usable samples of only several hundred observations are not unusual. The reasons for and consequences of such large data loss need to be addressed.

Other problems resulting from self-selection into a sample were identified 25 years ago by Saunders (1964) but have been largely ignored. A particularly troubling problem may be a school district's willingness to get involved in programs like DEEP and a student's desire to take economics courses, as recognized by Peterson. An analysis similar to that in Richard Murnane et al.'s (1985) study of the relative quality of public vs. Catholic schools may be useful in the analysis of the effect of one type of special program or project vs. another.

In the absence of randomized experiments, one can always find selection problems at some point in the sampling process. James Heckman and V. Joseph Hotz (1989) provide specification tests for selecting among alternative nonexperimental estimators. In considering selectivity problems, however, researchers should not lose sight of the relevant population to which generalizations are to be made. For instance, why should we care if we cannot teach a subject to the uninterested and unwilling? We are always going to be teaching to selected individuals, so why shouldn't our experiments reflect the actual conditions under which we work? Why worry about what will never apply? There are many situations in education for which potential selection problems can be raised but for which there are no practical consequences.

VI. Conclusion

What we know about the effectiveness of high school courses, materials, and organizations aimed at advancing an understanding of economics is empirical. Learning that is registered for one treatment vs. another is measured on an ordinal scale. At best relative comparisons are possible; absolute scores have little meaning.

Typically standardized test scores have not shown any great difference between one method of instruction or program versus another. This may simply reflect a shortcoming of these instruments, however. Consideration of alternatives to the standard paper and pencil test instruments for assessing the value of economic education efforts is needed. A better conceptual basis also is needed to integrate the why, how, and what teachers teach with what motivates students to learn. With alternative measures of success and a better conceptual base we might be able to give new answers to Stigler's old question: "Why should people be economically literate?"

REFERENCES

Baumol, William and Highsmith, Robert, "Variables Affecting Success in Economic Education: Preliminary Findings from a New Data Base," *American Economic Review Proceedings*, May 1988, *78*, 257–62.

Bach, G. L. and Saunders, Phillip, "Economic Education: Aspirations and Achievements," *American Economic Review*, June 1965, *55*, 329–56.

Becker, William, "The Educational Process and Student Achievement Given Uncertainty in Measurement," *American Economic Review*, March 1982, *72*, 229–36.

______ **and Walstad, William,** *Econometric Modeling in Economic Education Research*, Boston: Kluwer-Nijhoff, 1987.

______ **and** ______, "Data Loss From Pretest to Posttest as a Sample Selection Problem," *Review of Economics and Statistics*, forthcoming 1990.

Beron, Kurt, "Joint Determination of Current Classroom Performance and Additional Economics Classes: A Binary/Continuous Model," *Journal of Economic Education*, forthcoming 1990.

Bosshardt, William and Watts, Michael, "Instructor Effects and Their Determinants in Pre-College Economic Education," *Journal of Economic Education*, forthcoming 1990.

Buckles, Stephen and Morton, John, "The Effects of Advanced Placement on College Introductory Economics Courses," *American Economic Review Proceedings*, May 1988, *78*, 263–68.

______**, Strom, Robert and Walstad, William,** "An Evaluation of a State Consumer and Economic Education Program: Implications for Effective Program Delivery," *Journal of Economic Education*, Spring 1984, *15*, 101–10.

______ **and Freeman, Vera,** "A Longitudinal Analysis of a Developmental Economics Education Program," *Journal of Economic Education*, Winter 1984, *15*, 5–10.

Chizmar, John and Halinski, Ronald, "Performance in the Basic Economics Test (BET) and 'Trade-Offs'," *Journal of Economic Education*, Winter 1983, *14*, 18–29.

Chizmar, John et al., "Give and Take, Economics Achievement, and Basic Skills Development," *Journal of Economic Education*, Spring 1985, *16*, 99–110.

Dawson, George, "Research in Economic Education at the Precollege Level," in Donald Wentworth et al., eds., *Perspectives on Economic Education*. New York: JCEE, 1977, 85–103.

______, "A Note on the Teaching of Economics Through American History," *Journal of Economic Education*, Winter 1982, *13*, 65–68.

Eberts, R. and Stone, J., "Teacher Unions and the Productivity of Public Schools," *Industrial and Labor Relations Review*, April 1987, *40*, 354–63.

Ferber, Marianne, Birnbaum, Bonnie and Green, Carole, "Gender Differences in Economic Knowledge: A Reevaluation of the Evidence," *Journal of Economic Education*, Spring 1983, *14*, 4–10.

Green, Gary and Jaquess, Sue, "The Effect of Part-Time Employment on Academic Achievement," *Journal of Educational Research*, July/August 1987, *80*, 325–29.

Grimes, Paul and Register, Charles, "Teachers' Unions and Student Achievement in High School Economics," *Journal of Economic Education*, forthcoming 1990.

Hansen, W. Lee, Kelley, Allen and Weisbrod, Burton, "Economic Efficiency and the Distribution of Benefits from College Instruction," *American Economic Review Proceedings*, May, 1970, *60*, 364–69.

Heath, Julia, "An Econometric Model of the Role of Gender in Economic Education," *American Economic Review Proceedings*, May 1989, *79*, 226–30.

Heckman, James and Hotz, V. Joseph, "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," NBER Working Paper No. 2861, February 1989, 1–28.

Highsmith, Robert, *A Survey of State Mandates for Economics Instruction*, New York: Joint Council on Economic Education, 1989.

______ **and Baumol, William,** "Education in Economics: Evidence on Determinants of Effectiveness," mimeo., November 1989.

Jackstadt, Stephen and Gootaert, Christian, "Gender, Gender Stereotyping, and Socioeconomic Background as Determinants of Economic Knowledge and Learning," *Journal of Economic Education*, Winter 1980, *12*, 34–40.

Lewis, Darrell, Dalgaard, Bruce and Boyer, Carol, "Cost-Effectiveness of Computer Assisted Economics Instruction," *American Economic Review Proceedings*, May 1985, *75*, 91–96.

Lillydahl, Jane, "Academic Achievement and Part-Time Employment of High School Students," *Journal of Economic Education*, forthcoming 1990.

Lopus, Jane, "Do Additional Expenditures Increase Achievement in the High School Economics Class?," *Journal of Economic Education*, forthcoming 1990.

Lumsden, Keith and Scott, Alex, "The Economics Student Reexamined: Male-Female Differences in Comprehension," *Journal of Economic Education*, Fall 1987, *18*, 365–75.

Lynch, Gerald, "The Effect of Teacher Coursework on Student Learning: Evidence from the TEL," *Journal of Economic Education*, forthcoming 1990.

Marlin, James, "State Mandated Economic Education, Teacher Attitudes, and Student Learning," *Journal of Economic Education*, forthcoming.

Martin, David and Bender, David, "'Trade-Offs,' Field Dependence/Independence, and Sex-based Economics Comprehension Differences," *Journal of Economic Education*, Winter 1985, *16*, 62–70.

Murnane, Richard, Newstead, Stuart and Olsen, Ronald, "Comparing Public and Private Schools: The Puzzling Role of Selectivity Bias," *Journal of Business & Economic*

Statistics, January 1985, *3*, 369–77.

Myatt, Anthony and Waddell, Charles, "An Approach to Testing the Effectiveness of the Teaching and Learning of Economics at High School," *Journal of Economic Education*, forthcoming 1990.

Peterson, Norris, "Self-Selection in the NAEE Database," presented at Western Economics Association meeting, Lake Tahoe, Nevada, June 1989.

Rhine, Sherrie, "The Effect of State Mandates on Student Performance," *American Economic Review Proceedings*, May 1989, *79*, 231–35.

Rosen, Sherwin, "Some Economics of Teaching," *Journal of Labor Economics*, October 1987, *5*, 561–75.

Saunders, Phillip, "The Effectiveness of *The American Economy* in Teaching Secondary School Teachers," *American Economic Review*, June 1964, *54*, 396–400.

Schober, Howard, "The Effects of Inservice Training on Participating Teachers and Students in Their Economics Classes," *Journal of Economic Education*, Fall 1984, *18*, 282–96.

Siegfried, John, "Male-Female Differences in Economic Education: A Survey," *Journal of Economic Education*, Spring 1979, *10*, 1–11.

______ **and Fels, Rendigs,** "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, *17*, 923–69.

______ **and Walstad, William,** "Research on Teaching College Economics," in P. Saunders and W. Walstad, eds., *The Principles of Economics Course: A Handbook for Instructors*, New York: McGraw-Hill, 1990, 270–96.

Soper, John and Walstad, William, *The Test of Economic Literacy: Examiner's Manual*, 2nd ed., New York: Joint Council on Economic Education, 1987.

Stigler, George, "The Case, If Any, for Economic Education," *Journal of Economic Education*, Spring 1970, *1*, 77–84.

Thorton, Daniel and Vredeveld, George, "In-Service Education and Its Effect on Secondary Students: A New Approach," *Journal of Economic Education*, Spring 1977, *8*, 93–99.

Walstad, William and Soper, John, "What is High School Economics? Factors Contributing to Student Achievement and Attitudes," *Journal of Economic Education*, Winter 1989, *20*, 23–38.

______ **and** ______, "A Report Card on the Economic Literacy of U.S. High School Students," *American Economic Review Proceedings*, May 1988, *79*, 251–56.

______ **and Watts, Michael,** "Teaching Economics in the Schools: A Review of Survey Findings," *Journal of Economic Education*, Spring 1985, *16*, 135–46.

Watts, Michael, "A Statewide Assessment of Precollege Economic Understanding and DEEP," *Journal of Economic Education*, Summer 1985, *16*, 225–37; "Errata," *JEE*, Fall 1985, *16*, 286.

______, "Student Gender and School District Differences Affecting the Stock and Flow of Economic Knowledge," *Review of Economics and Statistics*, August 1987, *69*, 561–66.

______, "Research on DEEP: The First 25 Years," in W. Walstad and J. Soper, eds., *Economics in the School Curriculum: The DEEP Experience*, Washington: National Education Association, forthcoming 1990.

Weaver, Andrew, Deaton, William and Reach, Sheryll, "The Effect of Economic Education Summer Institutes for Teachers on the Achievement of Their Students," *Journal of Educational Research*, May/June 1987, *80*, 296–99.

Wentworth, Donald, Hansen, W. Lee and Hawke, Sharryl, *Perspectives on Economic Education*. New York: Joint Council on Economic Education, 1977.

THE NEW ECONOMICS OF PERSONNEL: RATIONALE OR RATIONALIZATION?[†]

Firm Organization and the Economic Approach to Personnel Management

By DEBRA J. ARON*

There has been limited application of the large theoretical literature in the areas of agency and monitoring to personnel management issues. The research developing and testing implications of agency theory on the design of compensation contracts has focused on *managerial* labor contracts; the theory of monitoring has seen even fewer applications to the personnel literature. The premise of this paper is that while agency issues may be directly applicable to managerial labor, for *nonmanagerial* personnel the monitoring problem must be addressed preliminary to the design of incentive contracts. I first describe the role of commitment in monitoring problems. This is then applied to a discussion of the difference between penalties and rewards in motivating workers. The relative success of Japanese manufacturing techniques is then put in the context of an organizational attempt to solve the monitoring problem. I conclude by suggesting areas for future research.

I. Agency and Monitoring: Applications to Personnel Management

What differentiates economic issues in personnel management from other aspects of manufacturing or production is, of course, that human workers require some sort of incentives to perform in their jobs as their employers would like them to. When a worker's job input or effort is perfectly and freely observable and verifiable, incentive provision is not, at least in principle, a problem. The first-best level or quality of input can be implemented by simply paying the worker appropriately as a function of his input. The extensive literature on agency, or "principal-agent" problems, has arisen to address the question of how to motivate workers when input is not observable, but some stochastic signal of input, such as output, can be costlessly observed.

One reason that most attempts to apply the agency literature are restricted to managerial labor markets is that typical compensation contracts for top managers include features that appear qualitatively consistent with the predictions of principal-agent models; for example, stock options, bonuses that depend on firm performance, and equity ownership. These make the manager's compensation indirectly a function of his own productivity by making it a function of the firm's market value.

Perhaps a more important reason that the recent literature has focused on managerial labor markets is that a strict interpretation of the models requires that the measure of output be freely observable and verifiable. A unique and critical feature of managerial labor markets is that the firm's stock value is indeed such a measure. Moreover, the stock value is a credible signal because it represents the consensus opinion of experts in the field who have a personal financial stake in making a correct assessment of the manager's performance (see my 1988 paper).

Clearly, the relationship between the decisions of a CEO and the firm's stock value is a very noisy and indirect one. Since factory workers are much closer to the final physical product, one might expect that incentive contracting would be more readily applied in those markets. However, for such workers a signal of the market value of output is, in general, not freely observable. Even in the simplest case of piece workers, and even if quality were not an issue, someone must count, measure, or audit the worker's output.

[†] *Discussants*: Michael J. Piore, MIT; Clair Brown, University of California-Berkeley; Edward Lazear, University of Chicago.

*Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208. I thank Eitan Zemel and Larry Manheim for helpful discussions.

Devising a system of *observation*, or monitoring, generates a number of delicate issues. For example, one must confront the problem of measuring quality in such a way that the worker's incentives are not distorted; and one must provide incentives for the monitor (if the monitor is human rather than mechanical) to monitor honestly and diligently. How one compensates the worker once the observations are made may then be an agency issue, but in general the incentive contracts will not be independent of the monitoring technology.

II. The Commitment Problem

A fundamental problem that arises in models of optimal monitoring is illustrated by the following example. Suppose that in order to monitor the accuracy of a typist, an employer could announce that she planned to proofread each letter (or, alternatively, one-third of the letters, chosen at random) for typos. If a typo was found the typist would be penalized. Suppose this were enough to make the typist so diligent that errors would never be made. Then the employer would have the incentive never actually to proofread the letters; knowing this, the employee would not have an incentive to be diligent. In this example, the equilibrium breaks down because it is not subgame perfect to monitor, that is, it is not optimal *ex post*. The employer's inability to commit to monitor the employee precludes existence of the equilibrium in which the worker never shirks.

Under what circumstances is it reasonable to think of commitment to a monitoring intensity feasible? Perhaps one way to precommit to a monitoring strategy is to implement mechanical monitoring. However, if the worker cannot tell whether or not the system is monitoring him, and if there is a variable cost associated with automated monitoring, the employer will still have an incentive to disable the monitoring system, and the lack of commitment problem persists. Further, if the *ex ante* optimal monitoring strategy is stochastic, one must be able to make credible to the employee that the random monitoring is in effect without making observable *ex ante* when it is and is not operating.

Alternatively, one may attempt to design the production technology so that there is no marginal cost to monitoring. For example, if all of the costs of an automated proofreading system are fixed (programming, installing the system, etc.) then the employer (weakly) has no *ex post* incentive to disable it. Another, organizational, approach is described in Section IV.

III. Are Rewards More Effective than Punishment?

Edward Lazear (1989) suggests another issue relevant to monitoring and the economics of personnel. Although psychologists and practitioners in personnel management treat bonus schemes and penalty schemes as quite different incentive mechanisms, economists have no framework for making sense of the distinction. Whether a contingent compensation scheme is "framed" (to use the terminology of Daniel Kahneman and Amos Tversky; see their 1979 and 1981 papers) or advertised to workers as a bonus or penalty should be a matter of indifference; in principle, any payment scheme that rewards the worker for high output can be written as a scheme that penalizes the worker for low output, and the actual payment contingent on any given output level will be the same under each contract.

There have been some attempts to reconcile theory and observation. Stefano Fenoaltea (1984) takes a quasi-psychological, quasi-economic approach to this issue by assuming that employers have at their disposal more than financial incentives; they can also impose physical punishment. He posits that the threat of physical pain creates anxiety, and cites evidence that increased anxiety increases work effort, but impairs the performance of skilled tasks. Thus, in jobs that require care and/or skill or are vulnerable to sabotage, one would expect to observe incentive provision by ordinary (financial) rewards. In jobs that are purely physical (such as digging ditches), pain incentives may be more efficient. Fenoaltea argues that this model is applicable to modern labor markets in that, for workers on the margin of poverty, the threat of dismissal (a penalty) is tantamount to the threat of hunger.

Some empirical implications suggest themselves. As technology allows replacement of pure physical labor with capital, the incidence of jobs appropriate for penalties decreases. Thus, we should see a secular decline in "penalty" schemes over time, as well as fewer penalty schemes in countries or industries with greater technological advancement. In addition, aspects of even highly skilled jobs that require no skill, such as arriving to the job on time, would be good candidates for penalty incentives.

My paper with Pau Olivella (1989) derives similar empirical implications from a purely economic model. We argue that a bonus or penalty scheme can only be meaningfully differentiated by what the payment stream looks like in equilibrium. A bonus is defined as anything given in addition to the *customary* amount; a penalty is a decrease in the customary payment. Thus, whether a scheme should be viewed as a penalty or bonus depends on what is "customary," which in turn depends on how frequently the worker is monitored (or "tested" or "evaluated") in equilibrium and whether the worker is likely to be diligent or shirking when the monitoring occurs. Suppose, for example, that the worker is promised a base wage, plus an additional payment every time he is monitored and found to be diligent in his job. If the equilibrium monitoring scheme has the worker being continuously monitored and if the worker is diligent 90 percent of the time, the observed payments would appear as a penalty scheme, since only 10 percent of the time would the worker receive less than the base wage plus incentive payment. On the other hand, if the equilibrium monitoring scheme had the worker being monitored only 10 percent of the time and the worker was always diligent, this would appear as a bonus scheme: 90 percent of the time the worker would receive only the base salary. Drawing on the work of Olivella (1989), it is shown that when commitment to a monitoring strategy is possible, the workers will (almost) always be diligent and will face penalties if they shirk. When commitment is infeasible, the optimal payment structure depends on whether the monitoring technology permits false positives. In the absence of false positives the workers will again face penalties if found shirking. However, when false positives are possible, there will be considerable shirking by workers in equilibrium, and they will get a bonus when found to be diligent.

The applicability of such a model depends on whether one can meaningfully interpret the crucial features of the monitoring technology (ability to commit and existence of false positives) in the context of real employment situations. In order for it to be impossible to incorrectly accuse a worker of shirking it must be very clear what the task is, how it is to be performed, and it must be possible to clearly define all dimensions of acceptable output. This becomes a more difficult task if quality is an important dimension of output. In such jobs one would expect to see bonus schemes. On the other hand, in jobs that require only purely physical labor it may be very clear by simple observation whether the worker shirked or not; in such tasks penalties are predicted.

As discussed earlier, it is difficult to find good examples of monitoring technologies in which the monitor can truly commit to a monitoring strategy, particularly if the strategy involves probabilistic monitoring. Nevertheless, the results suggest that it is in the interest of firms to develop monitoring technologies in which commitment is possible. Some American companies have developed systems that use computers to track the progress of manufactured equipment as it progresses through, for example, an assembly line. As each worker finishes his job, he enters his personal identification code and the code of the piece he worked on. In this way every step of assembly of each piece of equipment is associated with a particular worker, and defects can be traced back to individuals. It is interesting to note, however, that in the factories in which the author is aware that such a system is in place, the individual-specific information is not used for compensation purposes.

IV. Organizational Approaches to Monitoring in Japan

Recent history has seen a decline in importance of American manufacturing relative to the rest of the world, and Japan in particular.

It is often claimed that the Japanese success in manufacturing is not translatable to our country because it is based on cultural influences and social norms that do not prevail here. However, the critical feature of the highly touted Japanese manufacturing systems known as Just-in-Time (JIT) and Total Quality Control (TQC) may be unrelated to cultural or social differences, but rather to the particular personnel features inherent in those systems.

Although space does not permit a full description of these systems here, a simplified description follows (see Richard Schonberger, 1982). In a TQC factory the workers themselves, rather than a quality control department, are responsible for detecting defects. Rather than statistical sampling of finished or intermediate outputs, as in the West, each worker inspects each piece as it is received in the fabrication or assembly process. This is facilitated by the JIT system, in which parts are not fabricated in job shops and then passed in large lots to the next process, but proceed through each step of the fabrication process one (or few) at a time. This is similar to current practice in assembly in the West, but is an innovation in the fabrication stage of manufacturing. The crucial feature of this system is that, as each worker receives the next piece to work on, it is his responsibility to see that it has come to him with no defects. If it has, it is his responsibility to stop the line and bring the error to the attention of the previous worker. In this way the cause of the error is detected and corrected before a large lot is produced. Further, the workers who made the defective parts must do the rework to correct the errors, unlike typical Western factories in which there are separate rework crews. If the day's allotted work is not completed by the end of the shift, due to many reworks and errors, the workers may be required to stay late to finish the schedule. Further, workers are responsible for maintaining their own machines.

From the standpoint of the economics of personnel and incentives, the crucial feature of JIT/TQC is that the entire system appears to be designed to maximize observability and accountability of individual workers. The JIT system can be thought of as having a built-in monitoring system, in which each worker monitors the one before him. Each worker can be provided with incentives to report defects brought to him because either the defect prevents him from performing his task, or the defect will be traced back to him when later in the line the product does not function properly. Further, the fact that workers are responsible for their own machinery means that they cannot avoid accountability by blaming faulty equipment.

It should be noted that in order to implement these features the entire manufacturing process, including the design of the machines themselves, the physical configuration of the fabrication and assembly lines, the way in which the product moves through the factory, the pace of output, and the mix of products being produced must all be designed with these goals in mind. Two points are important here. First, when analyzing the enormous success of the Japanese in producing high-quality products, the critical feature is perhaps not the production process itself but rather the way that the production process facilitates improved monitoring and therefore incentives for workers. Second, it is important for economists interested in personnel issues, particularly as they relate to nonmanagerial workers, to take seriously the "institutional" features of the manufacturing system.

V. Areas of Further Research

The interaction between the manufacturing process and the incentives of workers has not been sufficiently studied in the economic literature (one interesting attempt is Masahiko Aoki, 1986). My paper with

Olivella argues, for example, that the JIT/TQC system is a way of endogenizing the ability to commit to a monitoring strategy. Further, a system in which the goal is 100 percent quality control and each defect represents shirking (which is not the case in the Western system of statistical sampling) decreases the scope for false positives to exist. Thus, the features crucial for determining whether bonuses or penalties are optimal and whether they are actually observed in equilibrium may be seen as part of the manufacturing system, and these features differ importantly across countries and industries.

To the extent that observability is built into Japanese manufacturing systems, and to some extent in Western systems, one would expect incentives to improve for workers; this could explain why Japanese workers are believed to be more conscientious than those in the West. I have here only argued that the systems differ in observability of workers. The final step in implementing incentives is translating the information from the monitoring system into a compensation scheme.

It is not at all clear how this is done in practice. It was noted above that, curiously, even in modern American factories in which a great deal of worker-specific information is available, this information is not used directly to determine the pay of the individual workers. It is not sufficient to argue that union rules prevent such use; this response begs the question of why the unions would want to implement such rules. Further, simple pay-for-performance schemes are apparently eschewed in nonunion firms as well (Fred Foulkes, 1980). It may be that the difficulty in distinguishing high- from low-quality work, or in the terminology of the monitoring literature, the likelihood of false positives, may make piece-rate incentive plans untenable.

Nor is it clear how observability translates into compensation in Japanese firms. It seems that formal records are not kept on the number of times an individual employee must rework a part, for example, or how often he stops the line when he detects a defect. As alluded to above, workers as a group may be penalized if there are many errors on a given day by requiring the shift to remain until the work is complete. But when it appears that the technology admits and even encourages individual evaluation, our simple models do not explain why individual rather than (or in addition to) group penalties are not instituted. Clearly, more subtle incentives are at work, and further research is required before we can understand why firms adopt such systems and why workers respond to them.

REFERENCES

Aoki, Masahiko, "Horizontal vs. Vertical Information Structure of the Firm," *American Economic Review*, December 1986, *76*, 971–83.

Aron, Debra, "Corporate Spinoffs in an Agency Framework," CMSEMS Discussion Paper No. 792, Northwestern University, June 1988.

______ **and Olivella, Pau,** "Bonuses vs. Penalties: An Economic Approach," mimeo., Northwestern University, December 1989.

Fenoaltea, Stefano, "Slavery and Supervision in Comparative Perspective: A Model," *Journal of Economic History*, September 1984, *44*, 635–68.

Foulkes, Fred K., *Personnel Policies in Large Nonunion Companies*, Englewood Cliffs: Prentice-Hall, 1980.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

Lazear, Edward P., "Economic Approaches to the Psychology of Organizations," Hoover Institution Working Paper E-89-15, May 1989.

Olivella, Pau, "Information Structure and the Incentive to Monitor: A Non-Cooperative Approach," unpublished doctoral dissertation, Northwestern University, June 1989.

Schonberger, Richard J., *Japanese Manufacturing Techniques*, New York: Free Press, 1982.

Tversky, Amos and Kahneman, Daniel, "The Framing of Decisions and the Psychology of Choice," *Science*, 1981, *211*, 453–58.

Who Bosses Whom? The Intensity of Supervision and the Discipline of Labor

By DAVID M. GORDON*

> The first person is called an employer and the second an employee. Calling the employer the boss is a custom derived from the fact that the "boss" specifies the particular task. One could have called the employee the boss because he orders the employer to pay him a specific sum if he wants services performed. But words are words.
>
> [Armen Alchian and William Allen, 1969, p. 320]

There now appears to be considerable convergence of interest between "modern" neoclassical and recent neo-Marxian analyses of the labor process, with both traditions paying increasing attention to the organization of production. (See for example, Francis Green, 1988.)

Despite this new attention to "command" structures governing production in the capitalist firm, however, remarkably few economists have yet analyzed one central element in that command structure—the "intensity of supervision," the ratio within the firm's hierarchy of supervisory to production-worker inputs. Legions of workers hang from supervisory rungs on the U.S. occupational ladder—agents of the "bosses," peering over the shoulders of their subordinates, poking and prodding, cajoling and commanding, managing and manipulating those below them in the production hierarchy. What do these supervisors do? Where do they sit and fit in relation to the more familiar categories of labor and management, employee and employer, worker and boss?

This paper reports on some initial explorations of the determinants of the intensity of supervision. The first section defines the intensity of supervision and presents summary data on trends and variations. The second section derives some skeletal analytic expectations about the determinants of supervisory intensity, highlighting conflicting expectations between efficiency-wage and labor-discipline models. The third section briefly reports results on some econometric tests of those analytic expectations for the United States both across manufacturing industries and over time for the private nonagricultural economy.[1] A final section comments on implications for further research.

I. Definitions and Trends

Following the lead of much of the recent literature on "shirking" and labor effort, we might compactly define supervisory inputs as those wage-and-salary employees of a firm with considerable if not primary responsibility for monitoring the labor effort of those below them in the firm's hierarchy, especially including the labor effort of production workers.

The closest approximation to measurement of these supervisory inputs available on a consistent and continuous basis for the United States comes from the Bureau of Labor Statistics (BLS) decomposition of total private nonagricultural employment into two exhaustive and mutually exclusive categories: one group of "production and related workers in manufacturing and mining, construction workers in construc-

---

*Department of Economics, Graduate Faculty, New School for Social Research, 65 Fifth Ave., New York, NY 10003. Effusive thanks to Subhash Gupta, David Howell, and Bruce Pietrykowski for their help in mobilizing some of the data for this exploration.

[1] Space limitations preclude reporting more than a few highlights of these specifications and results. A longer version of this paper, with detailed results and with a Data Appendix, is available from the author.

tion, and nonsupervisory employees in private service-producing industries" (U.S. Department of Labor, 1989, p. 143); and the residual excluded category of nonproduction/nonconstruction/supervisory employees. I shall label the ratio of nonproduction to production employees, defined by these BLS categories, as the NonProduction Index of Supervisory Intensity (s_{NP}). By this measure, there were 24 nonproduction employees for every 100 production workers in the United States in 1988.

Despite the obvious appeal of this measure, many included in the nonproduction category may have little or nothing to do with direct employee supervision. Two other more delimited definitions of the intensity of supervision are available for the United States on an intermittent, discontinuous basis. 1) It is possible to determine those occupations that are rated in the *Dictionary of Occupational Titles* to be proximately involved in directly supervisory activities and those that are directly supervised, permitting an alternative index of supervisory intensity. In 1985, by this measure, there were 15.5 employees in supervisory occupations for every 100 employees not in those occupations. 2) At least for the purposes of much more explicit focus on the point batallions of the supervisory army, we can concentrate on workers in manufacturing in the decennial Census 3-digit category of "foremen, not elsewhere classified (nec)." In 1985 in manufacturing, there were 5.1 foremen for every 100 workers not in supervisory occupations.

Figure 1 traces the NonProduction Index of Supervisory Intensity, highlighting its dramatic increase from .144 in 1951 to .237 in 1988, an increase of roughly two-thirds.

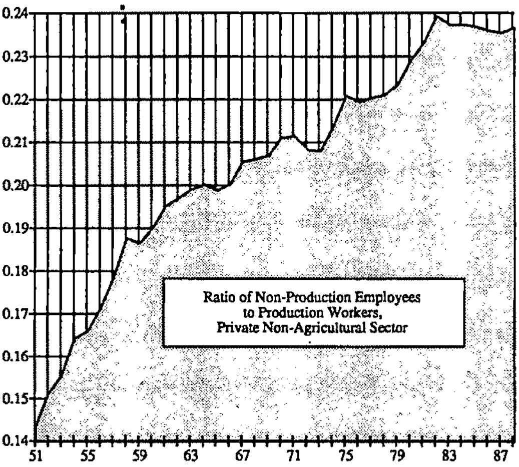


FIGURE 1. THE "INTENSITY OF SUPERVISION" IN THE UNITED STATES, 1951–88

TABLE 1—THE INTENSITY OF SUPERVISION IN THE UNITED STATES AND OTHER ADVANCED COUNTRIES, 1950–85

A. United States, Alternative Measures	1950	1960	1970	1980	1985
NonProd. Employees as % Pvt. NonAg. Employ.	12.3	16.0	17.4	18.7	19.2
Employees Supervisory Occs. as % Pvt. NonAg. Employ.	12.4	12.0	11.7	13.7	14.4
Foremen, nec as % Manufacturing Employ.	3.5	4.2	5.0	5.4	4.2
B. Selected Countries, Admin. & Managerial Employees as % NonAg. Employ.	1960	1970	1980	1985	
United States	6.6	8.7	11.4	11.7	
Japan	3.9	5.9	5.2	4.0	
West Germany	2.6	2.5	2.8	4.1	
Sweden	2.1	2.6	2.9	2.7	

Panel A of Table 1 displays data for the three different definitions of supervisory workers as a percent of total employment in the private nonagricultural sector (and in manufacturing for the category of foremen, nec). All three measures reflect significant increases in the intensity of supervision through most of the period of observation with two of the three continuing to rise in the early 1990s.

Panel B traces an overlapping category of all "administrative and managerial employees" as a percent of nonagricultural employment for the United States, Japan, West Germany, and Sweden, based on consistent translations of individual countries' occupational/industry censuses into standardized definitions. 1) For each of the years presented, proportionate employment in administrative and managerial employees was much higher in the United States than in the other three advanced economies. 2) Employment in this category grew much more rapidly as a share of private nonagricultural employment in the United States than in any of the others. Why are the legions of supervisory employees so vast and relatively growing in the United States?

II. Analytic Expectations

According to almost all of the recent literature highlighting variable labor effort, supervisory inputs are necessary to monitor both the intensity of labor services provided by production workers and the effectiveness of monitoring activities by their immediate supervisors. (See Carl Shapiro and Joseph Stiglitz, 1984; and Samuel Bowles, 1985, for examples.) This allows us to write a standard "effort" or "extraction" function:

$$l^* = l^*(w^*, s), \tag{1}$$

where l^* is labor intensity, w^* is the wage premium paid to workers to elicit work effort, and s is the intensity of supervision. Following Bowles, the firm chooses the optimal intensity of supervision which satisfies the first-order conditions $(l_s^*/l_{w^*}^*) = w_s$, where w_s is the hourly cost of a supervisory input, or the supervisory wage. At a minimum, following this logic, we might write a functional expression for the determinants of supervisory intensity as

$$s = s(w^*, w_s, Z), \tag{2}$$

where Z represents a vector of other factors affecting the labor extraction function $l^*(\cdot)$. In order to begin investigating the determinants of supervisory intensity, such a skeletal model would simply require fleshing out the vector Z and stipulating sign expectations about the first partials of the supervisory-intensity function, s_{w^*}, s_{w_s}, and s_Z.

At this point in the investigation, however, we confront a clear difference of emphasis between models emphasizing "efficiency wages" and those emphasizing "labor discipline." In many (though not necessarily all) efficiency-wage models, wage incentives are thought to constitute a sufficient inducement for workers to augment their labor effort *without complementary supervision*. In labor-discipline models, by contrast, wage incentives can extract labor effort *only if complemented by supervision*; if workers were not observed in their work, given conflicts of interest between employers and production employees, they would have no incentive to increase their labor effort even in return for a higher wage payment.

This leads to opposite expectations for the sign of the cross derivative between the intensity of supervision and the wage premium inducing labor effort. In many efficiency-wage models, the cross partial within equation (1) is normally postulated to be negative: $l^*_{w^* \cdot s} < 0$. (See, for example, Jonathan Leonard, 1987, equation [4], p. S142.) In labor discipline models, by contrast, the wage premium and supervisory intensity are stipulated as having a positive cross partial: $l^*_{w^* \cdot s} > 0$. (See, for example, Bowles, p. 23 and notes 14, 15.) Thus, in analyses of the intensity of supervision, we might reasonably distinguish between the two approaches by empirical evidence on the sign of s_{w^*} in (2): for efficiency-wage models, we expect $s_{w^*} < 0$, while for labor-discipline models, we expect $s_{w^*} > 0$.

III. Econometric Explorations

A. *Interindustry Variations*

In the absence of direct micro data on firms' supervisory efforts, cross-sectional data on industries may come closest to providing an appropriate arena for testing these analytic expectations about the determinants of supervision. I report here on cross-sectional analysis of sixty-eight 3-digit manufacturing industries for which relevant data were available on a continuous basis between 1958 and 1981. Given the data available, it was possible to test for five kinds of possible determinants of the intensity of supervision: the effects of variation in capacity utilization; in wage premia or rents, measured by variations in the "cost of job loss"; the probability of dismissal from the job; relative worker independence, proxied by the quit rate and union density; and the relative importance of other factor inputs, encompassing both capital intensity and materials intensity.

TABLE 2—SELECTED DETERMINANTS OF INTENSITY OF SUPERVISION ACROSS MANUFACTURING INDUSTRIES

	Coefficients		
	Cost of Job Loss	Union Density	Adj. R^2
1966	1.55[a]	−0.002[b]	0.360
1970	1.60[a]	−0.002[b]	0.296
1973	0.59	−0.003[a]	0.376
1975	1.93[b]	−0.005[a]	0.368
1979	0.50	−0.001	0.424

[a] Significant at 1 percent on two-tailed tests.
[b] Significant at 5 percent on two-tailed tests.

In the first set of explorations, I tested for consistency among the three alternative indices of supervisory intensity initially defined in Section I by regressing each of those dependent variables on the same estimating equation for 1970. Despite considerable differences among these alternative dependent variables, the labor-discipline model receives strong and relatively robust support in all three equations. Of special interest, the coefficient on the cost of job loss is positive and significant in all three equations, supporting expectations derived from the labor-discipline model about the effect of wage premia.

The NonProduction Index of Supervisory Intensity is by far the easiest of these measures to use, since it is available on a continuous and consistent basis for the entire (private nonfarm) economy. Heartened from this first experiment that models of variable labor effort appear to be as useful in explaining variations in s_{NP} as for the other two more explicit measures, I further estimated equations with the NonProduction Index of Supervisory Intensity as the dependent variable for business-cycle troughs and peaks from 1966 through 1979.

Table 2 summarizes some of the highlights of these estimations, reporting on coefficients and significance for measures of wage premia (the cost of job loss) and worker independence (union density), as well as the adjusted R^2 for the equations, for these five business-cycle turning points. Two results highlighted in the table seem most relevant for the purposes of this brief report:

1) The efficiency-wage model receives no support. In none of the five regressions is the coefficient on the wage-premium variable negative and significant. It is positive and significant (at least at 5 percent) in three of the five equations.

2) The usefulness of the labor-discipline perspective appears to decline over time. By the 1979 peak, neither of the two variables in Table 2 have significant coefficients. This appears at least partly to reflect some convergence late in the period in structures of supervision across industries; this may indicate, in turn, increasingly general reliance across industries, regardless of its effectiveness at the margin, on supervisory inputs as part of a more general "management offensive against labor" (see my study with Bowles and Thomas Weisskopf, 1990, chs. 5–6).

B. *Variations Over Time*

I next report on results of a time-series analysis of the NonProduction Index of Supervisory Intensity for the private nonfarm sector for the postwar period from 1951 through 1987. The model estimated is derived from the same basic model as for the cross-sectional estimations and contains roughly the same independent variables.

All independent variables have their expected signs and are statistically significant, further reinforcing impressions that labor-effort models may be useful for explaining variations in supervisory intensity. The coefficient on the cost of job loss is positive and significant (at 1 percent), further confirming expectations of the labor-discipline model and confounding those derived from the efficiency-wage approach. The model explains nearly all of annual variations in the NonProduction Index of Supervisory Intensity, with an adjusted R^2 of 0.991, without reliance on either a linear time trend or an autoregressive error-term adjustment.

IV. Conclusions

Who bosses whom? The relatively more harmonious emphases of "efficiency-wage" approaches and the relatively more conflictual accents of labor-discipline models project different expectations about the processes generating variations in the intensity of supervision.

The results reported here provide strong (but entirely provisional) support for the labor-discipline model. It would appear that bosses need supervisors in many situations because they cannot induce desired levels of labor effort without them—no matter how tasty the wage-premia carrots they may dangle in front of their workers.

But much more work needs to be pursued along many directions: A broader range of possible determinants of supervisory intensity needs to be considered in cross-sectional analyses and a broader spectrum of industries needs to be examined. Time-series analyses should be conducted not merely for the aggregate (private nonfarm) economy but also for disaggregated sectors and industries. Perhaps most critically, we need urgently to pursue cross-national comparisons: Why are supervisory armies relatively so vast in the United States? What role (if any) has this (apparently) conflictual style of labor management played in the eroding global competitiveness of the United States?

I have tried in this short paper to provide a case for taking supervisors seriously. There are millions of them. They apparently serve important economic functions. We should learn much more about them.

REFERENCES

Alchian, Armen A. and Allen, William R., *Exchange and Production: Theory in Use*, Belmont: Wadsworth, 1969.

Bowles, Samuel, "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models," *American Economic Review*, March 1985, *75*, 16–36.

______, **Gordon, David M. and Weisskopf, Thomas E.,** *After the Waste Land: A democratic economics for the year 2000*, Armonk: M. E. Sharpe, forthcoming 1990.

Green, Francis, "Neoclassical and Marxian Conceptions of Production," *Cambridge Journal of Economics*, September 1988, *12*, 299–312.

Leonard, Jonathan S., "Carrots and Sticks: Pay, Supervision, and Turnover," *Journal of Labor Economics*, October 1987, Part 2, *5*, S136–52.

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings*, November 1989, *36*.

Sticky Stories: Economic Explanations of Employment and Wage Rigidity

By SANFORD M. JACOBY AND DANIEL J. B. MITCHELL*

There was a time when "institutional" labor economists and economic theorists literally parted company over methodology. Specifically, in the late 1940s labor economists formed their own Industrial Relations Research Association (IRRA) as a rump group of the AEA. A notable feature of the IRRA was that its membership was drawn from both academics and labor-management practitioners. The idea was that practitioners, who knew the facts of the labor market, would inform academic researchers.

Knowledge of the stylized facts of the labor market may once have been disdained by mainstream economists, especially if these observations seemed to contradict theoretical expectations. At the same time, rigorous analytical and empirical approaches were often omitted from the institutionalists' tool kit. But in recent years there has developed a "new economics of personnel" (NEP) that bases its analysis on certain potentially uncomfortable facts, yet brings to bear a high level of rigor. Some (not all!) of the older doyens of labor market institutionalism have even taken pleasure in the "new generation of exceedingly competent labor economists" (Clark Kerr, 1983, p. 21). An important question, therefore, is whether the gap between institutional knowledge and theoretical analysis has now been bridged. We are encouraged by the bridging attempt, but argue that a gap remains due in part to a tendency for the stylized facts to shift.

I. Time and Space

The NEP offers efficiency explanations for modern American employment practices such as career labor markets, wage rigidity, and seniority-based layoffs. Its reasoning is functional; the practices must be efficient or else firms pursuing alternative policies would become dominant. The explanations proffered are derived from analysis of rational actors in what often seems to be a timeless, placeless steady state. Unfortunately, that mode of analysis cannot tell us what specifically is modern or American about current U.S. employment practices. This void would not be a problem were there but negligible national or historical variation in employer policy. However, the evidence shows such variation to be substantial.

American employment practices have rarely matched the textbook model of a spot labor market. But, over the course of this century, those practices went through a transition. Employment relationships became markedly more stable, while internal labor markets and seniority-layoff systems proliferated (see Jacoby, 1983). Similarly, nominal wages in the United States never were completely flexible. But, during the 1910s and 1920s, nominal wage cuts were more prevalent (and wage sensitivity to labor market conditions was greater) than in the post-World War II period (see Mitchell, 1985).

Europe and Japan did not go through the same transitions. The relative reliance on layoffs, nominal wage change, and real wage change varies from country to country. These time-and-space departures from U.S. arrangements suggest that rationalizations based on current American practices cannot be the basis of universal theory.

Confronted with the specificity of their stylized facts, NEP theorists have three responses. One is to argue that observed variations are transitory disequilibria; eventually employers will shift from less to more efficient employment practices. Oliver Williamson (1985) makes precisely this argument for the American historical case, infer-

---

*Anderson Graduate School of Management, UCLA, Los Angeles, CA 90024.

ring that present practices displaced earlier ones because of superior efficiency. But the inference depends on other assumptions. One is that market forces are always strong enough to weed out inefficiency. A second is that there is one right way to manage, that is, that multiple efficiency points do not exist.

Comparative observation suggests that multiple equilibria within a market setting are possible. On the one hand, critics of "Eurosclerosis" in the 1980s pointed to the American model of labor market "flexibility" (easy layoffs) as proof that European long-term job security was undesirable. At the same time, American management experts journeyed to Japan to marvel at long-term employment commitment and the ability it gave management to move workers to needed tasks.

A third response is to argue that any departure from current U.S. practices such as wage rigidity or career ladders must be due to variations in the incentives (for example, risk aversion, monitoring costs) that make those practices efficient. But historical research shows that, in the past, these incentives at least needed the push of social forces and historical accident (unions, wars, the Great Depression) to put now-existing practices into place. At the international level, economic rationality and technology (from which incentive structures should spring) are similar in modern industrial nations. So economic incentives can only be part of the story of historical and national variation in employer policies. The NEP either abstracts from such observations or, less often, stretches economic theory to endogenize them. A more fruitful approach would be to examine the interplay between institutions, incentives, and market outcomes, even if it means building partial models and acknowledging a limited ability to forecast.

II. Norms of Fairness

The institutionalists of the 1940s put great emphasis on notions of fairness in the labor market. For example, pattern bargaining in the union sector was seen as based on worker views of comparative equity. In contrast, NEP analysts would probably prefer to focus on reducing transactions costs in bargaining as a rational explanation for patterns. Yet there is other evidence that norms of fairness are important in labor markets. Moreover, these norms often seem to involve nominal—not real—measures.

TABLE 1—FORMAT OF REPORTED MAJOR PRIVATE UNION WAGE ADJUSTMENTS

	1959	1969	1979	1989
Nominal	69	83	80	68
Percent	12	12	19	25
Mixed[a]	15	2	1	1
Zero	2	0	0	6
Other	3	3	[b]	0

Source: Current Wage Developments, June issues of year shown.

Note: Percent in each category. Details need not sum to 100 due to rounding.

[a] Combined nominal and percentage description.

[b] Less than 0.5 percent.

Consider, for example, the annoying (to economists) tendency for union wage settlements to be reported in cents-per-hour terms rather than as percentage adjustments. Table 1 shows the proportion of private wage adjustments reported in the June issues of *Current Wage Developments* over the past three decades. Despite intervening episodes of significant inflation and wage guidelines and controls that emphasized percentages, the tenacity of wage setters in clinging to nominal expression of their settlements is noteworthy. The use of nominal expression suggests that when intersettlement comparisons are made for equity purposes, base wages matter little. That is, if two units receive 30 cent increases, workers in both feel that equity has been achieved even if one unit has a higher average wage (and therefore receives a smaller percentage increase). The tendency of escalator formulas to be specified in nominal terms (typically a cents-per-hour increase per index *point* increase) is a symptom of the nominal orientation. Employees apparently view wage relationships differently than do economists.

Survey data suggest a public aversion to nominal wage cutting. (Daniel Kahneman

TABLE 2—DISTRIBUTION OF EMPLOYEES UNDER MAJOR UNION WAGE SETTLEMENTS IN THE CONSTRUCTION INDUSTRY

	1980–81	1982	1983–84	1985–88
Decrease	0	5	13	3
Zero	0	10	41	26
0.1–1.9	7	3	16	16
2.0–3.9		10	8	27
4.0–5.9		16	7	20
6.0–7.9		17	6	7
8.0–9.9	15	19	5	
10.0–11.9	21	11	3	
12.0–13.9	17	4		
14.0 & up	39	5		
Mean Adjustment	13.4	6.5	1.1	2.3

Source: Current Wage Developments, various issues.
Note: Multiyear figures are weighted by annual number of employees covered. Percent in each category. Details need not add to 100 due to rounding.

et al., 1986.) While such aversion might be explained as a reflection of a rationally based implicit contract by NEP theorists, the fact that the concern is focused on the nominal wage (not the real wage) creates obvious analytical problems. Table 2 presents the distribution of union wage settlements in the construction industry during the 1980s, a period when the industry was especially prone to "concession bargaining."[1] The distribution shows a decided downward shift in mean adjustments beginning in 1982. Yet the distribution bunches at exactly the (nominal) zero level. While some workers did receive wage cuts, zero was clearly a resistance point for pay adjustments.

J. M. Keynes (1936, p. 14) was quite willing to accept fairness as a basis for apparent money illusion in the labor market; it formed a major element in the *General Theory*. In his view, decentralized wage setting led to a worker focus on comparative (nominal) wages, even if price movements eroded purchasing power. Would today's NEP analysts be willing to buy such an explanation? And even if he or she did, what would be done with the observation that the degree of real vs. nominal wage rigidity seems to vary across countries? The Keynesian vs. classical debate over European unemployment in the 1980s suggests that worker norms of fairness can change over time and vary across countries.

III. Changing Times

Although NEP provides a rationale for various anomalous facts first pointed out by labor economists of the 1940s and 1950s, it appears that this theoretical success has come just at the time when the facts themselves are changing rapidly. Labor markets are moving away from the rigidities captured by those facts. Having rationalized a difference between labor and product markets, NEP analysts must now consider the possibility of greater similarity across these markets than existed in the past.

Impelling the move away from rigid forms of employment are cost pressures arising from a more open and competitive economy. One response to those pressures was the wave of permanent layoffs that swept American industry during the early to mid-1980s, eliminating many previously secure jobs. Disproportionately affected were older, blue-collar, unionized, manufacturing workers. Managers have not been exempt from such cost-cutting moves.

A related development was the shift of employment during the last decade to smaller firms and establishments, where compensation is lower and personnel practices are less formalized. To compete, large establishments are using part-time employees or are contracting work out to the burgeoning business service industries that supply temporary workers and other business support services. While not all of the work being contracted out is done at lower wage and benefit costs, the bulk of business services are provided by lower-status workers whose pay is less than that of regular, in-house employees. Generally, full-timers are paid more than comparable part-timers, with the latter predominantly made up of female clerical and service workers.

---

[1] Construction is presented because most contracts in the industry do not feature escalator clauses or lump sum bonuses that obscure actually experienced pay adjustments. However, distributions similar to those of Table 2 characterized the entire union sector in the 1980s.

Part-time, temporary, and business service workers comprise the contingent labor force, each segment of which has grown more rapidly since 1980 than the labor force as a whole. In addition to cost factors, use of contingents has been spurred by an unstable and less predictable business environment, which results from more variable product demand, shorter product life cycles, and flexible exchange rates. Employers are trying to keep a lean core of employees with job security and to smooth out demand fluctuations by using contingent workers who lack such security.

Tenure data offer a way of tracing these shifts in employment stickiness. Table 3 shows that between 1963 and 1981, job tenure for men declined but there was little or no change for women or mature workers. Because workers can change jobs without changing employers, a more accurate gauge of employment stickiness is tenure with a given employer. This measure has been used by the Bureau of Labor Statistics since 1983, unfortunately breaking the older series. As shown in Table 4, between 1983 and 1987, groups experiencing sharp tenure declines included mature men (previously a stable group); male managers and machine operators; and female clerical and service workers. Those outcomes reflect an expansion in contingent labor and shifts of employment away from stable jobs in large firms.

TABLE 3—MEDIAN TENURE OF U.S. EMPLOYEES ON CURRENT JOB (YEARS)

	All Males	Males 45–54	All Females	Females 45–54
1963	5.7	11.4	3.0	6.1
1966	5.2	11.5	2.8	5.7
1968	4.8	11.3	2.4	5.1
1973	4.6	11.5	2.8	5.9
1978	4.5	11.0	2.6	5.9
1981	4.0	11.0	2.5	5.9

Source: U.S. Bureau of Labor Statistics, *Special Labor Force Reports*, various issues, and Bulletin no. 2162 (1983). Figures are for January.

TABLE 4—MEDIAN TENURE OF U.S. EMPLOYEES WITH CURRENT EMPLOYER (YEARS)

	Females		Males	
	1983	1987	1983	1987
All	3.7	3.6	5.1	5.0
25–34	3.2	3.1	3.8	3.7
45–54	6.9	7.3	13.2	12.3
Managers	4.6	4.9	7.5	6.9
Prof'l.	4.7	5.0	6.1	6.6
Sales	2.5	2.5	4.7	4.4
Clerical	3.9	3.7	[b]	[b]
Service[a]	2.8	2.5	[b]	[b]
Operators	[b]	[b]	5.6	5.1
Craft	[b]	[b]	5.5	5.5

Source: U.S. Bureau of Labor Statistics, "Job Tenure: 1983, 1987," unpublished tables. Figures are for January.

[a] Excluding protective service workers.

[b] Not one of the five largest occupations for that sex.

Wages also are becoming less rigid and more sensitive to firm performance, despite the pressures emanating from norms of fairness. In 1986, profit sharing in medium-to-large firms stood at 22 percent of full-time employees. Lump sum bonuses, that hardly existed before the 1980s, were in the contracts of 51 percent of private, unionized, nonconstruction workers under major agreements in late 1988. There is some evidence (not definitive as yet) that these bonuses will prove more variable than wage rates and that they are a response (like contingent employment) to demand uncertainty (Christopher Erickson and Andrea Ichino, 1989). As variable pay methods spread, the kind of wage stickiness associated with long-term employment is likely to be reduced.

Why has the NEP literature been slow to recognize and acknowledge these changes? As in other fields, scholars tend to learn a given set of facts and stick with them. This tendency was characteristic of experts on Eastern Europe (who until recently insisted that nothing was changing in that part of the world) as well as U.S. industrial relations scholars (who were slow to recognize the pace of union sector shrinkage during the 1980s).

While slow response is not unique to NEP, it is nonetheless true that the NEP's static approach to theory is not well-suited for predicting outcomes of dynamic relationships, such as those between firms and their environments. As in other branches of economics, the NEP tends to give more emphasis to internal theoretical consistency than to external consistency with a welter of changeable facts.

IV. Where Do We Go from Here?

The NEP analysts have demonstrated something important; almost any behavior can be rationalized within the neoclassical framework. While the usual fallback is to use empirical testing to sort out the true explanation, such testing in practice rarely sorts out competing views definitely. Starting from stylized facts, if nothing else, prevents one from saying silly things about nonexistent phenomena. But neither institutionalists nor NEP theorists can base their observations on a set of stylized facts, and ignore trends which are changing those facts, without risking obsolescence.

Some physicists reportedly feel they are on the verge of a Theory of Everything, a unified theory that will encompass all known forces. But it is unlikely that such a theory will arrive for labor market analysts anytime soon. What we can hope for are informed but modest partial theories and explanations.

REFERENCES

Erickson, Christopher and Ichino, Andrea, "Lump Sum Bonuses in Union Contracts," working paper, MIT, 1989.

Jacoby, Sanford M., "Industrial Labor Mobility in Historical Perspective," *Industrial Relations*, Spring 1983, *22*, 261–82.

Kahneman, Daniel, Knetch, Jack L. and Thaler, Richard, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, *76*, 728–41.

Kerr, Clark, "A Perspective on Industrial-Relations Research—Thirty-Six Years Later," in Barbara D. Dennis, ed., *Proceedings of the Thirty-Sixth Annual Meeting*, Madison: IRRA, 1984.

Keynes, John Maynard, *The General Theory of Employment, Interest, and Money*, New York: Harcourt, Brace, 1936.

Mitchell, Daniel J. B., "Wage Flexibility: Then and Now," *Industrial Relations*, Spring 1985, *24*, 266–79.

Williamson, Oliver E., *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*, New York: Free Press, 1985.

AGGREGATE ASSET PRICING[†]

Asset Prices under Habit Formation and Catching up with the Joneses

By ANDREW B. ABEL*

This paper introduces a utility function that nests three classes of utility functions: 1) time-separable utility functions; 2) "catching up with the Joneses" utility functions that depend on the consumer's level of consumption relative to the lagged cross-sectional average level of consumption; and 3) utility functions that display habit formation. Incorporating this utility function into a Lucas (1978) asset pricing model allows calculation of closed-form solutions for the prices of stocks, bills and consols under the assumption that consumption growth is i.i.d. Then equilibrium asset prices are used to examine the equity premium puzzle.

I. The Utility Function

At time t, each consumer chooses the level of consumption, c_t, to maximize $E_t(U_t)$ where $E_t\{\ \}$ is the conditional expectation operator at time t and the utility function is given by

$$(1)\qquad U_t \equiv \sum_{j=0}^{\infty} \beta^j u(c_{t+j}, \nu_{t+j})$$

where ν_{t+j} is a preference parameter. Suppose that the preference parameter ν_t is specified as

$$(2)\qquad \nu_t \equiv \left[c_{t-1}^{D} C_{t-1}^{1-D}\right]^{\gamma} \quad \gamma \geq 0 \quad \text{and} \quad D \geq 0$$

where c_{t-1} is the consumer's own consumption in period $t-1$ and C_{t-1} is aggregate consumption per capita in period $t-1$. If $\gamma = 0$, then $\nu_t \equiv 1$ and the utility function in (1) is time separable. If $\gamma > 0$ and $D = 0$, the parameter ν_t depends only on the lagged level of aggregate consumption per capita. This formation is the relative consumption model or "catching up with the Joneses."[1] Finally, if $\gamma > 0$ and $D = 1$, the parameter ν_t depends only on the consumer's own past consumption. This formulation is the habit formation model.

Consider the effects on utility of a change in an individual's consumption at date t, holding aggregate consumption unchanged. Substituting (2) into (1) and then differentiating with respect to c_t yields

$$(3)\qquad \partial U_t/\partial c_t = u_c(c_t, \nu_t) + \beta u_\nu(c_{t+1}, \nu_{t+1})\,\gamma D \nu_{t+1}/c_t.$$

Suppose that the period utility function $u(c_t, \nu_t)$ has the following isoelastic form

$$(4)\qquad u(c_t, \nu_t) = [c_t/\nu_t]^{1-\alpha}/(1-\alpha), \quad \alpha > 0.$$

When $\gamma = 0$, the utility function in (4) is the standard constant relative risk-aversion utility function and α is the coefficient of relative risk aversion. More generally, utility depends on the level of consumption relative to some endogenous time-varying benchmark ν_t.[2] Under the isoelastic utility function in (4), the expression for $\partial U_t/\partial c_t$ in (3) becomes

---

[†] *Discussants*: Phillippe Weil, Harvard University; Narayana Kocherlakota, Northwestern University; Stanley Zin, Carnegie Mellon University.

*Department of Finance, Wharton School of the University of Pennsylvania, Philadelphia, PA 19104. I thank Mike Perigo for helpful discussion and excellent research assistance.

[1] The phrase "catching up with the Joneses," rather than "keeping up with the Joneses," reflects the assumption that consumers care about the *lagged* value of aggregate consumption. The April 1989 version (p. 10) of Jordi Gali (1989), but not the September revision, examines the utility function $u(c_t, C_t) = [1/(2-\beta-\gamma)]c_t^{1-\beta}(c_t/C_t)^{-(1-\gamma)}$ and shows that when $\beta = 1$, asset pricing will be equivalent to an economy without consumption externalities and with log utility.

$$(5)\quad \partial U_t/\partial c_t = \left[1-\beta\gamma D(c_{t+1}/c_t)^{1-\alpha}(\nu_t/\nu_{t+1})^{1-\alpha}\right] \times (c_t/\nu_t)^{1-\alpha}(1/c_t).$$

II. Equilibrium

Let y_t be the amount of the perishable consumption good per capita produced by the capital stock. In equilibrium, all output is consumed in the period in which it is produced, as in Lucas. Because all consumers are identical, $c_t = C_t = y_t$ in every period. Now let $x_{t+1} \equiv y_{t+1}/y_t$ be the gross growth rate of output. Because $c_t = C_t = y_t$, it follows that $c_{t+1}/c_t = C_{t+1}/C_t = x_{t+1}$. Therefore, equation (2) implies that $\nu_{t+1}/\nu_t = x_t^{\gamma}$ which allows us to rewrite (5) as

$$(6)\qquad \partial U_t/\partial c_t = H_{t+1}\nu_t^{\alpha-1}c_t^{-\alpha}$$

where $H_{t+1} \equiv 1-\beta\gamma D x_{t+1}^{1-\alpha}x_t^{-\gamma(1-\alpha)}$.

Note that $H_{t+1} \equiv 1$ if $\gamma D = 0$, which is the case for both time-separable and relative consumption preferences.[3]

III. Asset Pricing

To calculate asset prices, let us examine a consumer who considers purchasing an asset in period t and then selling it in period $t+1$. If asset prices are in equilibrium, this pair of transactions does not affect expected discounted utility. Suppose that a consumer reduces c_t by 1 unit, purchases an asset with a gross rate of return R_{t+1}, sells the asset in period $t+1$, and increases c_{t+1} by R_{t+1} units. The equilibrium rate of return R_{t+1} must satisfy

$$(7)\quad E_t\{-(\partial U_t/\partial c_t) + \beta R_{t+1}(\partial U_{t+1}/\partial c_{t+1})\} = 0.$$

Equation (7) can be rewritten as

$$(8)\quad E_t\{\beta R_{t+1}(\partial U_{t+1}/\partial c_{t+1}) / E_t\{\partial U_t/\partial c_t\}\} = 1.$$

Equation (8) is the familiar result that the conditional expectation of the product of the intertemporal marginal rate of substitution and the gross rate of return equals one.[4] We can obtain an expression for $(\partial U_{t+1}/\partial c_{t+1})/E_t\{\partial U_t/\partial c_t\}$ using equation (6) to divide $\partial U_{t+1}/\partial c_{t+1}$ by $E_t\{\partial U_t/\partial c_t\}$ to obtain

$$(9)\quad (\partial U_{t+1}/\partial c_{t+1})/E_t\{\partial U_t/\partial c_t\} = \left[H_{t+2}/E_t\{H_{t+1}\}\right]x_t^{\gamma(\alpha-1)}x_{t+1}^{-\alpha}.$$

IV. The Price of Risky Capital

Let p_t^S be the exdividend price of a share of stock in period t, which is a claim to a unit of risky capital. The rate of return on stock is $R_{t+1}^S \equiv (p_{t+1}^S + y_{t+1})/p_t^S$. Let $w_t \equiv p_t^S/y_t$ be the price-dividend ratio. Therefore, $p_t^S = w_t y_t$ and $p_{t+1}^S = w_{t+1}y_{t+1}$ so that

$$(10)\qquad R_{t+1}^S = (1+w_{t+1})x_{t+1}/w_t.$$

Substituting (10) into (8) yields

$$(11)\quad w_t = \beta E_t\{(1+w_{t+1})x_{t+1} \times (\partial U_{t+1}/\partial c_{t+1})/E_t\{\partial U_t/\partial c_t\}\}.$$

---

[2] George Constantinides (1988), Jerome Detemple (1989), John Heaton (1989), and Suresh Sundaresan (1989) also examine asset prices in the presence of habit formation. James Nason (1988) includes a time-varying benchmark level of consumption that differs from habit formation in that it is independent of an individual consumer's own consumption.

[3] A sufficient condition for $\partial U_t/\partial c_t > 0$ when $\gamma = D = 1$ (habit formation) is $1 + \ln\beta/\ln(\max\{x\}/\min\{x\}) < \alpha < 1 + \ln\beta/\ln(\min\{x\}/\max\{x\})$. For $\beta = 0.99$ and the 2-point distribution in Table 1, the sufficient condition is $0.858 < \alpha < 1.142$.

[4] In the conventional time-separable formulation of this problem, $\partial U_t/\partial c_t$ is known as of time t, and hence $E_t\{\partial U_t/\partial c_t\}$ on the left-hand side of (8) equals $\partial U_t/\partial c_t$.

V. Bills and Consols

A one-period riskless bill can be purchased in period t at a price of s_t; in period $t+1$, the bill is worth 1 unit of consumption. The gross rate of return on the bill is $R^B_{t+1} \equiv 1/s_t$. Substituting $1/s_t$ for the rate of return in (8) yields

$$s_t = \beta E_t\{(\partial U_{t+1}/\partial c_{t+1}) / E_t\{\partial U_t/\partial c_t\}\}. \tag{12}$$

A consol bond, that pays one unit of consumption in each period, can be purchased at an excoupon price p^C_t in period t. In period $t+1$, the consol pays a coupon worth one unit of consumption and then sells at a price of p^C_{t+1}. The one-period rate of return on the consol is $R^C_{t+1} \equiv (1+p^C_{t+1})/p^C_t$. Substituting R^C_{t+1} into (8) yields

$$p^C_t = \beta E_t\{(1+p^C_{t+1})(\partial U_{t+1}/\partial c_{t+1}) / E_t\{\partial U_t/\partial c_t\}\}. \tag{13}$$

VI. I.I.D. Consumption Growth

Suppose that consumption growth x_{t+1} is i.i.d. over time. In this case, we can obtain explicit solutions for the prices of stock, bills, and consols. The price-dividend ratio w_t is

$$w_t = A x^\theta_t / J_t, \tag{14}$$

where $\theta \equiv \gamma(\alpha - 1)$

$$A \equiv \beta E\{x^{1-\alpha}\}\left[1-\beta\gamma D E\{x^{(1-\alpha)(1-\gamma)}\}\right] / \left[1-\beta E\{x^{(1-\alpha)(1-\gamma)}\}\right]$$

$$J_t \equiv E_t\{H_{t+1}\} \equiv 1 - \beta\gamma D E\{x^{1-\alpha}\} x^\theta_t.$$

The price of a one-period riskless bill is

$$s_t = q\beta x^\theta_t / J_t, \tag{15}$$

where

$$q \equiv E\{x^{-\alpha}\} - \beta\gamma D E\{x^{1-\alpha}\} E\{x^{\theta-\alpha}\}$$

and the price of a consol is

$$p^C_t = Q x^\theta_t / J_t, \tag{16}$$

where $Q \equiv \beta q / \left[1 - \beta E\{x^{\theta-\alpha}\}\right]$.

Given a distribution for x, the moments of x can be calculated and the three asset prices are easily calculated. For time-separable preferences ($\gamma = 0$) and relative consumption ($\gamma > 0$; $D = 0$), we can obtain closed-form solutions (in terms of preference parameters and the moments of x) for the unconditional expected returns $E\{R^S\}$, $E\{R^B\}$ and $E\{R^C\}$:

$$E\{R^S\} = E\{x^{-\theta}\} \times \left[E\{x\} + A\,E\{x^{1+\theta}\}\right]/A \tag{17}$$

$$E\{R^B\} = E\{x^{-\theta}\}/\beta q \tag{18}$$

$$E\{R^C\} = E\{x^{-\theta}\}\left[1 + Q E\{x^\theta\}\right]/Q. \tag{19}$$

Under habit formation, unconditional expected returns can be calculated numerically using the asset prices in (14)–(16).

VII. The Equity Premium

Rajnish Mehra and Edward Prescott (1985) report that from 1889 to 1978 in the United States, the average annual real rate of return on short-term bills was 0.80 percent and the average annual real rate of return on stocks was 6.98 percent. Thus the average equity premium was 618 basis points. They calibrated an asset pricing model with time-separable isoelastic utility to see whether the model could deliver unconditional rates of return close to the historical average rates of return on stocks and bills. They used a 2-point Markov process for consumption growth with $E\{x_t\} = 1.018$, $\text{Var}\{x_t\} = (0.036)^2$, and correlation $(x_t, x_{t-1}) = -0.14$. For values of the preference parameters that Mehra and Prescott deemed reasonable, the model could not produce more than a 35 basis point equity premium ($E\{R^S\} - E\{R^B\}$) when the expected riskless rate, $E\{R^B\}$, was less than or equal to 4

TABLE 1—UNCONDITIONAL EXPECTED RETURNS
$\beta = 0.99$; $E\{x\} = 1.018$; $\text{VAR}\{x\} = (0.036)^2$

α	Stocks	Bills	Consols
A. Time-separable preferences ($\gamma = 0$)			
0.5	1.93	1.87	1.87
	[1.93]	[1.87]	[1.87]
1.0	2.83	2.70	2.70
	[2.83]	[2.70]	[2.70]
6.0	10.34	9.52	9.52
	[10.33]	[9.51]	[9.51]
10.0	14.22	12.85	12.85
	[14.13]	[12.72]	[12.72]
B. Relative consumption ($\gamma = 1$; $D = 0$)			
0.5	2.80	2.76	2.73
	[2.80]	[2.76]	[2.73]
1.0	2.83	2.70	2.70
	[2.83]	[2.70]	[2.70]
6.0	6.70	2.07	5.84
	[6.72]	[2.06]	[5.86]
10.0	14.73	1.59	13.16
	[14.95]	[1.55]	[13.32]
C. Habit formation ($\gamma = 1$; $D = 1$)			
0.86	33.56	4.53	35.25
0.94	6.83	3.48	7.44
1.00	2.83	2.70	2.70
1.06	8.43	1.93	7.40
1.14	38.28	0.93	35.16

percent per year. This result is the equity premium puzzle.

Table 1 reports the unconditional expected rates of return on stocks, bills, and consols under the assumption that x_t is i.i.d., $E\{x\} = 1.018$ and $\text{Var}\{x\} = (0.036)^2$. For time-separable and relative consumption preferences, two unconditional expected returns are reported in each cell: the first is calculated under a 2-point i.i.d. distribution; the second, shown in brackets, is calculated under a lognormal distribution for x.

Panel A of Table 1, which reports the unconditional expected rates of return under time-separable preferences, displays the equity premium puzzle. Although $E\{R^S\}$ increases as α increases from 0.5 to 10.0, $E\{R^B\}$ also increases. The equity premium, $E\{R^S\} - E\{R^B\}$, does not come anywhere close to the 600-point historical average. Incidentally, the unconditional expected rates of return of bills and consols are exactly equal under time-separable preferences.

Panel B reports the unconditional expected rates of return in the relative consumption model. For $\alpha = 6$, the equity premium is 463 basis points and the unconditional riskless rate is 2.07 percent per year. Although the unconditional expected returns on stocks and bills are much closer to their historical averages, the conditional expected rates of return (not reported in the table) vary too much. For the 2-point distribution for x, the standard deviation of $E_t\{R^B_{t+1}\}$ is 17.87 percent when $\alpha = 6$. This unrealistic implication of the model poses a challenge for future research.

Panels A and B report unconditional rates of return for a lognormal distribution with $E\{x\} = 1.018$ and $\text{Var}\{x\} = (0.036)^2$. For the parameter values reported, it makes no substantial difference for expected returns whether the growth rate is lognormal or has a 2-point distribution.

Panel C presents the unconditional expected rates of return under habit formation. The expected rates of return on both long-lived assets (stocks and consols) are extremely sensitive to the value of α. Under logarithmic utility ($\alpha = 1$), the expected rates of return are the same as under time-separable preferences and relative consumption. However, with $\alpha = 1.14$, the expected rates of return on stocks and consols are both greater than 35 percent.

Further research using the utility function introduced in this paper will explore the implications of other settings for the parameters γ and D. For instance, if D is between zero and one, the utility function would contain elements of both catching up with the Joneses as well as habit formation. Also the assumption of i.i.d. consumption growth rates can be relaxed, and asset prices can then be analyzed numerically.

REFERENCES

Constantinides, George M., "Habit Formation: A Resolution of the Equity Premium Puzzle," University of Chicago, October 1988.

Detemple, Jerome, "General Equilibrium with Time Complementarity: Consumption, Interest Rates, and the Equity Premium," Graduate School of Business, Columbia University, September 1989.

Gali, Jordi, "Keeping up with the Joneses

Consumption Externalities, Portfolio Choice and Asset Prices," mimeo., Graduate School of Business, Columbia University, September 1989.

Heaton, John, "An Empirical Investigation of Asset Pricing with Temporally Dependent Preference Specifications," MIT, December 1989.

Lucas, Robert E., Jr., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, *46*, 1429–45.

Mehra, Rajnish and Prescott, Edward C., "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, March 1985, *15*, 145–61.

Nason, James, "The Equity Premium and Time-Varying Risk Behavior," Finance and Economics Discussion Paper Series, No. 11, Federal Reserve Board, February 1988.

Sundaresan, Suresh M., "Intertemporally Dependent Preferences and the Volatility of Consumption and Wealth," *Review of Financial Studies*, 1989, *2*, 73–89.

Measuring the Persistence of Expected Returns

By JOHN Y. CAMPBELL*

A great deal of recent research has documented the fact that rational expectations of real returns on long-term financial assets move systematically through time. Most of this work concentrates on the *variability* of expected returns, but the *persistence* or serial correlation of expected returns is also an important issue. If expected returns follow a persistent time-series process, then movements in expected returns will have a large impact on asset prices; prices are much less sensitive to transitory fluctuations in expected returns. Thus any attempt to explain the variability of asset prices or returns requires information on the persistence of movements in expected returns.

Information on persistence can also guide the search for an economic explanation of movements in expected returns. Variable expected returns need to be accounted for by variable intertemporal marginal rates of substitution, either those of a representative agent in a standard finance-theoretic model, or those of optimizing agents in a model with "noise traders." The time-series properties of expected returns can be used to restrict the time-series properties of the relevant intertemporal marginal rate of substitution.

Here I estimate the persistence of expected returns in the U.S. stock market, and use the estimates to account for the variability of unexpected stock returns. I begin with a standard regression model which forecasts monthly stock returns with a modest R^2 statistic of less than 7 percent. I incorporate this model into a vector autoregression that describes the evolution through time of the forecasting variables, and thus the expected stock return. I then calculate the effect on the stock price of an innovation in the expected return; this effect is large because movements in the expected return are persistent. Finally, I use the system to decompose the variance of the unexpected stock return into three components: the variance of changes in rational expectations of future returns, the variance of changes in rational expectations of future dividends, and a covariance term. The variance of changes in expectations of future returns and the covariance term are always important components of the overall variance of stock returns.

Another way to express this result is that the overall variance of stock returns is always greater than the variance of news about cash flows. Short-term predictability of returns can increase the variance of unexpected returns to a surprising degree. The findings here suggest that a satisfactory explanation of stock market volatility must simultaneously explain the short-term predictability of stock returns.

I. Expected Returns and Unexpected Returns

In order to characterize the relation between changing expected returns and unexpected returns, I will use the log-linear "dividend-ratio model" developed in my work with Robert Shiller (1988a; b). This model is an appropriate framework because it relates asset prices to both expected returns and expected future cash flows. The model says that, to a first-order Taylor approximation,

$$(1) \qquad h_{t+1} \approx k + \delta_t - \rho\delta_{t+1} + \Delta d_{t+1},$$

where h_{t+1} is the log holding-period return on a stock held from the end of period t to the end of period $t+1$, δ_t is the log dividend-price ratio $d_t - p_t$, d_t is the log dividend paid during period t, p_t is the log stock price at the end of period t, and ρ and k are parameters of linearization. The parameter ρ

---

*Woodrow Wilson School, Princeton University, Princeton NJ 08544. I am grateful to the National Science Foundation and the Sloan Foundation for financial support, and to Robert Shiller for exceptional assistance.

is the average ratio of the stock price to the sum of the stock price and the dividend, a number a little smaller than one.[1] Equation (1) says that the return on stock is high if the dividend-price ratio is high when the stock is purchased, if dividend growth occurs during the holding period, and if the dividend-price ratio falls during the holding period.

Equation (1) can be thought of as a difference equation relating δ_t to δ_{t+1}, Δd_{t+1} and h_{t+1}. Solving forward, and imposing the terminal condition that $\lim_{j\to\infty}\rho^j\delta_{t+j}=0$, one obtains

$$(2)\quad \delta_t = \sum_{j=1}^{\infty} \rho^{j-1}\left[h_{t+j} - \Delta d_{t+j}\right] - k/(1-\rho).$$

This equation says that a high dividend-price ratio today must generate high future returns unless dividend growth is low in the future. It is important to note that all the variables in (2) are measured *ex post*; (2) has been obtained *only* by the linear approximation of h_t and the imposition of a condition that δ_{t+j} does not explode as j increases.

However (2) also holds *ex ante*. If one takes rational expectations of equation (2), conditional on information available at the end of time period t, the left-hand side is unchanged since δ_t is in the information set, and the right-hand side becomes an *expected* discounted value. Using the *ex ante* version of (2) to substitute δ_t and δ_{t+1} out of (1), I obtain

$$(3)\quad h_{t+1} - E_t h_{t+1} = (E_{t+1} - E_t)\sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j} - (E_{t+1} - E_t)\sum_{j=1}^{\infty}\rho^{j} h_{t+1+j}$$

or in more compact notation,

$$v_{t+1} = \eta_{d,t+1} - \eta_{h,t+1}.$$

Equation (3) makes the central point of this paper. If expectations are internally consistent, then an unanticipated capital gain on the stock must be associated with either an increase in expected future dividends $\eta_{d,t+1}$ or a fall in expected future returns $-\eta_{h,t+1}$. To see why this is so, consider an asset with fixed dividends whose price rises. This asset's dividend yield is now lower, so its returns must be lower at some point in the future unless the asset price follows an explosive upward path which I have ruled out by assumption.

Equation (3) can be used to decompose unexpected stock returns into the two components $\eta_{d,t+1}$ and $-\eta_{h,t+1}$. In what follows, I shall refer to the former as "news about future dividends," and to the latter as "news about future expected returns." This should not be taken to imply one-way causality from expected dividends and returns to prices; in general, these variables are all determined simultaneously.[2]

The importance of persistence can be most easily appreciated by working out in detail the special case in which the expected return follows a first-order univariate autoregressive, or AR(1) process with coefficient ϕ. I define u_{t+1} to be the innovation at time $t+1$ in the one-period-ahead expected return, $u_{t+1} \equiv (E_{t+1} - E_t)h_{t+2}$. In the AR(1) case, u_{t+1} satisfies

$$(4)\quad E_{t+1}h_{t+2} = \phi E_t h_{t+1} + u_{t+1}.$$

It is important to note that equation (4) does not restrict the size of the market's information set. In particular, there is no presumption that the relevant information set contains only the history of past asset returns. It is quite possible that a very large number of variables is useful in forecasting the asset return over the next period; equation (4) merely restricts the way in which the next period's forecast is related to past forecasts.

---

[1] Equation (1) and the other formulas given here differ slightly from those given in my papers with Shiller because the notation here uses a different timing convention. In this paper I define the time t stock price and conditional expectation of future variables to be measured at the end of period t rather than the beginning of period t. This conforms with the more standard practice in the finance literature.

[2] Consider, for example, a model in which noise traders become bullish on stock, driving prices up and rational expectations of future returns down. In this model, the news is simply the innovation in the noise traders' demand for stock. Nonetheless (3) must still describe the expectations of any rational agent.

Equation (4) implies that $\eta_{h,t+1} = \rho u_{t+1}/(1-\rho\phi)$. Since ρ is a number very close to one, this means that a 1 percent increase in the expected return today is associated with a capital loss of about 2 percent if the AR coefficient is 0.5, a loss of about 4 percent if the AR coefficient is 0.75, and a loss of about 10 percent if the AR coefficient is 0.9. James Poterba and Lawrence Summers (1988) give a similar result.

One can also calculate the ratio of the variance of news about future expected returns to the total variance of unexpected returns. Equations (3) and (4) imply that for ρ close to one, $\mathrm{Var}(\eta_{h,t+1})/\mathrm{Var}(v_{t+1}) \approx ((1+\phi)/(1-\phi))(R^2/(1-R^2))$, where R^2 is the fraction of the variance of stock returns that is predictable. If R^2 is 0.025, then the share of news about future expected returns in the variance of unexpected returns is 0.08 for $\phi = 0.5$, 0.18 for $\phi = 0.75$, and a startling 0.49 for $\phi = 0.9$. These parameters are not unreasonable ones for monthly stock returns. An R^2 of 0.025 is quite modest, and a process with $\phi = 0.9$ has a half life of only a little more than 6 months. The lesson is that apparently trivial but persistent movements in expected returns can be a major force driving unexpected returns. The R^2 for one-period returns is thus an inadequate measure of the importance of expected return variation.

The assumption that the expected stock return follows a univariate AR(1) is highly restrictive. Fortunately, it is possible to generalize the previous discussion to handle the case where the expected return is one element of a vector autoregression (VAR). Shmuel Kandel and Robert Stambaugh (1989) also propose a VAR to describe stock returns.

First, I define a vector z_{t+1} that has k elements, the first of which is the stock return h_{t+1}. The other elements are other variables that are known to the market by the end of period $t+1$. Then I assume that the vector z_{t+1} follows a first-order VAR:

$$z_{t+1} = Az_t + w_{t+1}. \tag{5}$$

The assumption that the VAR is first-order is not restrictive, since a higher-order VAR can always be stacked into first-order (companion) form in the manner discussed in my paper with Shiller (1988a). The matrix A is known as the companion matrix of the VAR.

The first-order VAR generates simple multiperiod forecasts of returns. Define $e1$ to be a k-element vector, whose first element is 1 and whose other elements are all 0. This vector picks out the stock return from the variables of the VAR: thus the unexpected stock return $v_{t+1} = e1'w_{t+1}$. Expected future stock returns are given by $E_t h_{t+1+j} = e1'A^{j+1}z_t$. It follows that the discounted sum of revisions in forecast returns can be written as

$$\begin{aligned}\eta_{h,t+1} &= e1' \sum_{j=1}^{\infty} \rho^j A^j w_{t+1} \\ &= \lambda' w_{t+1},\end{aligned} \tag{6}$$

where λ' is defined to equal $\rho e1'(I-\rho A)^{-1}A$, a nonlinear function of the VAR coefficients. From equation (3), $\eta_{d,t+1}$ equals $(e1'+\lambda')w_{t+1}$. These expressions can be used to decompose the variance of the unexpected stock return, v_{t+1}, into the variance of the component that is due to expected return variation, $\eta_{h,t+1}$, the variance of the news about dividends, $\eta_{d,t+1}$, and a covariance term.

In the VAR context, there is no single measure of the persistence of expected returns. But one natural way to summarize persistence is by the variability of the innovation in the expected present value of future returns, relative to the variability of the innovation in the one-period-ahead expected return. Thus I define the VAR persistence measure P as

$$\begin{aligned}P &\equiv \sigma(\eta_{h,t+1})/\sigma(u_{t+1}) \\ &= \sigma(\lambda' w_{t+1})/\sigma(e1'Aw_{t+1}),\end{aligned} \tag{7}$$

where $\sigma(x)$ denotes the standard deviation of x. Another way to describe the statistic P is to say that a typical 1 percent innovation in the expected return will be associated with a P percent capital loss on the stock. In the univariate AR(1) case discussed earlier, $P = \rho/(1-\rho\phi) \approx 1/(1-\phi)$.

II. Application to the U.S. Stock Market

For the sake of comparability with previous work, I use a standard data set here. I study the behavior of the monthly value-weighted New York Stock Exchange Index, as reported by the Center for Research in Security Prices (CRSP) at the University of Chicago. The data set runs from 1926 to 1988, but I reserve the first year for lags so that my full sample period is 1927–88. I deflate the nominal return on the index using the Consumer Price Index reported in Ibbotson Associates (1989).

The forecasting variables I use for the stock return are the lagged stock return, the dividend-price ratio D/P, and the "relative bill rate" $RREL$, the difference between a short-term Treasury bill rate and its one-year backward moving average. The lagged stock return is included because the stock return forecasting equation will be one equation of a VAR system. The dividend-price ratio is included because the *ex ante* version of equation (2) shows that it will reflect any changes that may occur in future expected returns. (See also Eugene Fama and Kenneth French, 1988.) The ratio is measured as total dividends paid over the previous year, divided by the current stock price.

The relative bill rate is included because many authors have noted that the level of short-term interest rates helps to forecast stock returns. The short-term interest rate itself may be nonstationary over this sample period, so it needs to be stochastically detrended. The subtraction of a one-year moving average is a crude way to do this; the relative bill rate can also be written as a triangular moving average of changes in the short-term interest rate, so it is stationary in levels if the short rate is stationary in differences.[3] The short rate used is the one-month Treasury bill rate series from Ibbotson Associates.

One problem that arises when interest rate data are used is that the behavior of interest rates has changed over time. In particular, the Federal Reserve Board held interest rates almost constant for much of the period up to 1951, when a Federal Reserve Board-Treasury Accord allowed rates to move more freely. Accordingly, I split the 1926–88 sample at the end of 1951. This also allows a separate look at the data from the period around the Great Depression, which may behave differently from the postwar data (Myung Jig Kim et al., 1989).

The stock return forecasting equation estimated over the period 1927–88 is $h_{t+1} = 0.107h_t + 0.331(D/P)_t - 0.424RREL_t$. The equation has an R^2 of 0.024, and the forecasting variables are jointly significant at the 1.8 percent level using a heteroskedasticity-consistent test statistic. When D/P is regressed on the same forecasting variables, its own lag receives a coefficient of 0.963 while the other variables have almost zero coefficients. $RREL$ behaves in a similar fashion with an own lag coefficient of 0.669.

These estimates have quite striking implications for the variance decomposition of unexpected stock returns. The variance of news about future expected returns $\eta_{h,t+1}$ accounts for 0.28 of the total variance of unexpected stock returns, the variance of news about future dividends $\eta_{d,t+1}$ accounts for 0.37, while the covariance of these two terms accounts for the remaining 0.35. The covariance term is large because news about future returns and news about future dividends have a negative correlation of -0.53. When the stock market rises because of good news about future dividends, expected returns fall and the market rises even further.

Shocks to expected returns in this system can come from any of the three variables in the VAR. Each type of shock has a different persistence. The average persistence measure P is estimated to be 4.8, indicating that a typical 1 percent shock to the expected return is associated with a capital loss of almost 5 percent.

When the VAR system is estimated for the period 1927–51, the forecasting variables are jointly significant for stock returns at only the 18 percent level. The problem is partly that stock returns have a high variance that moves with the level of the dividend-price ratio; this means that heteroskedasticity correction greatly increases the standard error on this variable. Unsurprisingly, given the interest rate regime, the interest rate variable is a poor forecaster before 1952. Nonetheless, even in this period, the variance decomposition of unexpected returns attributes only 0.44 of the variance to the variance of news about dividends. The remainder is attributed to the variance of news about expected returns (0.19) or the covariance term (0.38).

---

[3] Another recently popular way to detrend the interest rate is to use the yield spread between interest rates of two different maturities. The relative bill rate has at least as much forecasting power for stock returns as the long-short yield spread, which is insignificant when it is added to the equations reported below.

In 1952–88 stock returns are more strongly forecastable; the forecasting system used here does *not* derive its predictive power from the Great Depression period. The returns equation has an R^2 of 6.5 percent, and both the dividend-price ratio and the relative bill rate are highly significant. Now 0.77 of the variance of unexpected stock returns is attributed to news about future expected returns, and only 0.13 to news about dividends. Once again the covariance term is positive because there is a negative correlation between news about dividends and news about expected returns.

The results presented here (and in greater detail in my 1990 paper) suggest that predictable time variation in stock returns is extremely important for understanding the volatility of the stock market. Predictable returns move quite persistently, and they tend to fall when expected dividends rise, amplifying the response of the stock market to news about future dividends. These facts pose a challenge to asset pricing theory.

REFERENCES

Campbell, John Y., "A Variance Decomposition for Stock Returns," NBER Working Paper No. 3246, January 1990.

——— **and Shiller, Robert J.,** (1988a) "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, Fall 1988, *1* 195–228.

——— **and** ———, (1988b), "Stock Prices, Earnings, and Expected Dividends," *Journal of Finance*, July 1988, *43*, 661–76.

Fama, Eugene F. and French, Kenneth R., "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, October 1988, *22*, 3–25.

Ibbotson Associates, *Stocks, Bonds, Bills, and Inflation: 1989 Yearbook*, Chicago, 1989.

Kandel, Shmuel and Stambaugh, Robert F., "Modelling Expected Stock Returns for Long and Short Horizons," unpublished paper, Wharton School, University of Pennsylvania, July 1989.

Kim, Myung Jig, Nelson, Charles R. and Startz, Richard, "Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence," unpublished paper, University of Washington, May 1989.

Poterba, James M. and Summers, Lawrence H., "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, October 1988, *22*, 27–59.

Evaluating Empirical Tests of Asset Pricing Models: Alternative Interpretations

By STEPHEN G. CECCHETTI AND NELSON C. MARK*

Recent tests of asset pricing models have proceeded in one of two ways. The first method, implemented by Lars Hansen and Kenneth Singleton (1982), begins by estimating parameters using various moment restrictions implied by an economic model, and then tests a set of overidentifying restrictions. In the second, due to Rajnish Mehra and Edward Prescott (1985) and known as calibration, a model is *tested* by seeing if it can exactly reproduce various moments of asset price data for given parameter values. Both of these procedures have shortcomings. The usual practice of testing overidentifying restrictions both involves asking whether a model can match aspects of the data that may be of little economic importance and fails to isolate those characteristics of the data that the model does not capture. Model calibration, on the other hand, does not take account of the statistical properties of both the sample moments trying to be matched and the model implied values to which they are being compared.

This paper describes an alternative strategy for testing asset pricing models that combines features from these two earlier approaches. Following Hansen and Singleton, we use statistical inference techniques to establish formal criteria that a successful model must satisfy. But in the spirit of Mehra and Prescott, we attempt to isolate moments of the data that we feel are of economic importance.

Our method has four natural steps. First we specify an economic model that allows us to characterize the evolution of asset prices. This requires a parameterization of both preferences and the stochastic process driving the endowment in the model. Following our earlier work with Pok-sang Lam (1990a, b), we assume that the utility function is of the constant relative risk aversion class and that the endowment can be accurately characterized by the Markov switching model pioneered by James Hamilton (1989).

In the second step, we choose a set of moments of the asset price data that we use to test the model. Clearly, any simple, parsimoniously parameterized model cannot match more than a small set of properties of the data. So instead of requiring that a model be capable of emulating all of the features of the data, we proceed to choose a small number of moments that are of particular economic interest. Following the literature, we examine the first and second moments of the equity premium and the risk free rate, as well as the variance ratio statistics computed from equity returns first presented by James Poterba and Lawrence Summers (1988).

The third step is estimation. In order to test whether the model is capable of matching the moments specified in the second step, we need to estimate the moments, the parameters of the endowment process and their joint covariance matrix.

In the fourth and final step we test the model. Following Mehra-Prescott and the calibration literature, we do not estimate the preference parameters. Instead we fix the discount factor (β) and the coefficient of relative risk aversion (γ), and then ask whether the moments that are implied by the model are close to those computed from the asset price data.

The methodology is simple and can be applied in any setting where model calibration has previously been used. Since it allows the investigator to choose various moments of the data for study, the method allows us to identify the specific points where the

---

*Ohio State University, Columbus, OH 43210. Cecchetti is also with the NBER; he acknowledges financial support from the National Science Foundation.

model is failing. This can help direct investigations into new and better alternative models.

In our study of asset returns, we find that there exist (β, γ) pairs for which the moments chosen are not rejected at standard levels of statistical significance (i.e., the 5 percent level). But in order to achieve this result, we must allow the coefficient of relative risk aversion to exceed 20. We conclude by commenting on what we believe to be appropriate criteria for choosing γ.

I. The Model

This section summarizes both a model of asset pricing and the solution for returns based on the endowment process introduced by Hamilton. Details can be found in our (1990a) paper with Lam.

Begin by specifying preferences as

$$U(C_t) = (C_t^{1-\gamma} - 1)/(1-\gamma), \tag{1}$$

where $0 < \gamma < \infty$ is the coefficient of relative risk aversion and C_t is the level of per capita real consumption.

Using (1), the standard first-order conditions that describe the evolution of the prices of traded assets can be written as

$$P_t^e C_t^{-\gamma} = \beta E_t C_{t+1}^{-\gamma}(P_{t+1} + D_{t+1}), \tag{2}$$

$$P_t^f = E_t \beta (C_{t+1}/C_t)^{-\gamma}; \tag{3}$$

where P_t^e is the real price of the traded asset, or equity; P_t^f is the real price of the risk-free asset; D_t is the payoff or dividend from owning one unit of equity; β is the discount factor; and E_t is the mathematical expectation conditional on information at time t.

To obtain a closed-form solution to the functional equations (2) and (3), we must specify the stochastic process governing $\{C_t, D_t\}$. In our earlier paper with Lam (1990b), we show that both consumption and dividend growth have significant negative skewness and leptokurtosis that is well captured by Hamilton's Markov switching model. We assume that the log of C_t and D_t, c_t and d_t, are given by a bivariate random walk with a two-state Markov drift. The model can be written as

$$\begin{pmatrix} c_t \\ d_t \end{pmatrix} = \begin{pmatrix} c_{t-1} \\ d_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_0^c \\ \alpha_0^d \end{pmatrix} + \begin{pmatrix} \alpha_1^c \\ \alpha_1^d \end{pmatrix} S_{t-1} + \begin{pmatrix} \varepsilon_t^c \\ \varepsilon_t^d \end{pmatrix}, \tag{4}$$

where $(\varepsilon_t^c \varepsilon_t^d)'$ is jointly i.i.d. normal with mean zero and covariance matrix Σ, and S_t is a Markov random variable that takes on values of 0 or 1 with transition probabilities $\Pr[S_t = 1 | S_{t-1} = 1] = p$ and $\Pr[S_t = 0 | S_{t-1} = 0] = q$. The model requires estimation of the parameter vector $\mathbf{B} = (\alpha_0^c, \alpha_0^d, \alpha_1^c, \alpha_1^d, p, q, \Sigma)$.

The model can now be solved for the prices of the risky and the riskless assets as functions of the process parameters $\mathbf{B}$, the state S_t, and the preference parameters β and γ. The result can be written as

$$P_t^e = \lambda(\mathbf{B}, \beta, \gamma; S_t) D_t; \tag{5}$$

$$P_t^f = \theta(\mathbf{B}, \beta, \gamma; S_t). \tag{6}$$

The various moments of the equity return, the equity premium and the risk-free rate implied by the model can be computed directly from (5) and (6). In general, these can be written as $M(\mathbf{B}; \beta, \gamma)$. In the next section we choose to study the first and second moments of the equity premium and the risk-free rate, as well as the variance ratios computed from equity returns. The variance ratios are simple linear functions of the autocorrelations of the returns data and are all equal to one if returns are truly serially independent.

II. Estimation and Testing

In order to evaluate the performance of the model described in Section I, and to perform the desired test, we need to estimate the process parameters $\mathbf{B}$, and the vector of moments of interest, call these μ. Using annual data from 1889 to 1987, we have computed estimates of both the process pa-

TABLE 1:—SELECTED SAMPLE MOMENTS OF ASSET RETURNS

	Mean	St. Dev.
Risk-Free Rate (R^f, σ_f)	1.15	5.87
Risk Premium (R^p, σ_p)	6.03	19.15
Correlation (ρ_{fp})	−0.02	
	Variance Ratios	
Two-Year (VR_2)	0.99	
Three-Year (VR_3)	0.83	
Five-Year (VR_5)	0.82	
Ten-Year (VR_{10})	0.71	

TABLE 2:—MODEL EVALUATION USING VARIOUS SAMPLE MOMENTS

	Preference Parameters		
β	1.135	1.095	0.945
γ	14.00	21.75	30.00
Moments Implied by the Model, $M(\hat{\mathbf{B}}, \beta, \gamma)$			
First and Second Moments			
R^f	1.00	1.86	5.63
R^p	1.77	2.78	4.00
σ_f	9.86	13.71	17.38
σ_p	11.26	21.88	47.61
ρ_{fp}	0.02	0.02	0.01
Variance Ratio Statistics			
VR_2	0.89	0.71	0.74
VR_3	0.83	0.56	0.61
VR_5	0.78	0.42	0.50
VR_{10}	0.74	0.31	0.41
Marginal Significance Level for Tests			
$K(R^f, R^p)$	0.54	0.24	0.59
$K(R^f, R^p, \sigma_p, \sigma_f, \rho_{fp})$	0.00	0.55	0.24
$K(VR_2, \dots, VR_{10})$	0.12	0.84	0.20

rameters from data on total per capita consumption (including durables) and dividends ($\hat{\mathbf{B}}$), and a set of sample moments from asset return data ($\hat{\mu}$), along with an estimate of the covariance matrix of $(\hat{\mathbf{B}}\hat{\mu})'$. Possible sample moments to be considered are reported in Table 1. (While the table only reports the variance ratios at two-, three-, five- and ten-year horizons, all the tests reported below include statistics at all horizons from two to ten years.)

Table 1 reflects a number of well-known aspects of the returns data. While the risk-free rate averaged approximately 1 percent, the equity premium exceeded 6 percent. In addition, the standard deviation of the equity premium is more than three times that of the risk-free rate. Finally, the variance ratio statistics lie below 1.0, implying a small amount of negative serial correlation in equity returns.

Model performance is evaluated by examining the difference between the model's implied values for the moments of interest, $M(\hat{\mathbf{B}}; \beta, \gamma)$, and the estimates from the asset price data, $\hat{\mu}$. We wish to find out if $[M(\hat{\mathbf{B}}; \beta, \gamma) - \hat{\mu}]$ is small. The estimate of this difference is plagued by uncertainty that arises from the fact that both the moments of the data, μ, and the process parameters, $\mathbf{B}$, must be estimated. Taking account of both sources of uncertainty, we can compute the sampling distribution of $[M(\hat{\mathbf{B}}; \beta, \gamma) - \hat{\mu}]$ directly from the estimate of the covariance matrix of the vector $(\hat{\mathbf{B}}\hat{\mu})'$; call this $\hat{\Gamma}$.

The resulting test statistic is given by

$$(7) \quad K = \left[M(\hat{\mathbf{B}}; \beta, \gamma) - \hat{\mu}\right]' \hat{\Gamma}^{-1} \times \left[M(\hat{\mathbf{B}}; \beta, \gamma) - \hat{\mu}\right],$$

which is distributed as *chi*-squared with degrees of freedom equal to the number of moments being considered, under the null hypothesis that $M(\mathbf{B}; \beta, \gamma) = \mu$. (See our paper with Lam, 1990a, for details.) When K is small, as measured by a high p-value from the cumulative *chi*-squared distribution, we can conclude that the model's implied moments are close to the data.

Implementation of this procedure requires the choice of a (β, γ) pair. Table 2 presents calculations for three possibilities. The general result is that the model comes fairly close to the sample moments when considering either the variance ratio statistics, or the means of the equity premium and the risk free rate. But the second moments of the premium and the risk-free rate are slightly more difficult to match. In fact, the point estimates for the standard deviations, σ_f and σ_p, exceed the sample values by apparently large amounts. But because the standard errors about these standard deviations are quite large, the marginal significance levels of the tests including them remain high.

All of the examples in Table 2 involve values of γ that may seem large.[1] It is important to note that there exist no (β, γ) pairs with γ less than 15.0, for which even the simplest mean test does not reject the model at the 5 percent significance level. But if γ is allowed to exceed 20, then the model performs very well.

---

[1] In addition, we have included cases in which $\beta > 1.0$. Both Narayana Kocherlakota (forthcoming) and our

III. Discussion

Traditionally, researchers have restricted their attention to values of γ that are less than 10.0. But the methodology we employ to evaluate the intertemporal capital asset pricing model finds that the coefficient of relative risk aversion must lie outside of this range in order for the model to perform well. Given this evidence, we have two choices. We can either abandon this simple formulation of the model because it requires values of γ that are unacceptably large, or we can change the acceptable range for γ.

Our view is that it is useful to consider values for the coefficient of relative risk aversion that are larger than have been considered in the past. The purpose of modeling is, after all, to provide an organizing framework that helps us to make sense of the data. To this end, a simple model is always preferred to a more complex one. There is mounting evidence that the basic model presented in this paper provides a consistent picture when γ is large, but not when it is small. (See also Shmuel Kandel and Robert Stambaugh, 1988.) In short, the data seem to be telling us that γ is larger than previously thought.

It is interesting to note that introspection can also reveal values of γ that are larger than previously agreed upon. Inspired by a question first put forth by Kandel and Stambaugh, we recently asked members of the faculty of the Department of Economics at Ohio State and the attendees of the session at the AEA annual meeting where this paper was presented to make an estimate of their income and then to answer the following question: *Based on the flip of a fair coin, I will either pay you \$1000 or you will have to pay me \$1000. How much would you be willing to pay me to avoid this bet?* We obtained a total of 22 responses to this question from which we constructed estimates of the coefficient of relative risk aversion (using the constant relative risk aversion utility function given in equation (1)). The responses implied γ's ranging from 0 to 77, with an average of 12.7 and a standard deviation of 15.7. Although the size of the sample is small, this evidence suggests that concentrating on γ greater than 10 is not unreasonable.

REFERENCES

Cecchetti, Stephen G., Lam, Pok-sang and Mark, Nelson C., (1990a) "The Equity Premium and the Risk Free Rate: Matching the Moments," Department of Economics, Ohio State University, February 1990.

______, ______, and ______, (1990b) "Mean Reversion in Equilibrium Asset Prices," *American Economic Review*, June 1990, *80*, forthcoming.

Hamilton, James D., "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, March 1989, *57*, 357–84.

Hansen, Lars P. and Singleton, Kenneth J., "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, September 1982, *50*, 1269–86.

Kandel, Shmuel and Stambaugh, Robert F., "Modeling Expected Stock Returns for Long and Short Horizons," mimeo., Department of Finance, Wharton School, University of Pennsylvania, December 1988.

Kocherlakota, Narayana R., "A Note on the 'Discount' Factor in Growth Economies," *Journal of Monetary Economics*, forthcoming.

Mehra, Rajnish and Prescott, Edward C., "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, March 1985, *15*, 145–61.

Poterba, James M. and Summers, Lawrence H., "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, October 1988, *22*, 27–59.

1990a paper with Lam discuss why it is sensible to consider discount factors in excess of 1.0.

STOCK MARKET VOLATILITY[†]

Do Security Analysts Overreact?

By WERNER F. M. DE BONDT AND RICHARD H. THALER*

It has long been part of the conventional wisdom on Wall Street that financial markets "overreact." Both casual observation and academic research support this view. The October crashes of 1987 and 1989 reinforce the research by Robert Shiller and others that suggests that stock prices are too volatile. Also, Shiller's 1987 survey evidence reveals that investors were reacting to each other during these crashes, rather than to hard economic news. A similar conclusion is reached by Kenneth French and Richard Roll (1986) who find that prices are more volatile when markets are open than when they are closed.

Our own prior research (1985, 1987) argued that mean reversion in stock prices is evidence of overreaction. In our 1985 paper, we showed that stocks that were extreme "losers" over an initial three- to five-year period earned excess returns over the subsequent three to five years. In the 1987 paper, we showed that these excess returns cannot easily be attributed to changes in risk, tax effects, or the "small firm anomaly." Rather, we argued that the excess returns to losers might be explained by biased expectations of the future. We found that the earnings for losing firms had fallen precipitously during the formation period (while they were losing value), but then rebounded strongly over the next few years. Perhaps, we speculated, "the market" did not correctly anticipate this reversal in earnings. This hypothesis, of excessive pessimism about the future prospects of companies that had done poorly, was suggested by the work of Daniel Kahneman and Amos Tversky (1973). They found that people's intuitive forecasts have a tendency to overweight salient information such as recent news, and underweight less salient data such as long-term averages.

Of course, there are many reasons to be skeptical that actual investors (stock market professionals) are subject to the same biases as student subjects in laboratory experiments. Definitely, the market professionals are experts in their field, they have much at stake, and those who make systematic errors may be driven out of business. Therefore, we present here a study of the expectations of one important group of financial market professionals: security analysts who make periodic forecasts of individual company earnings. This is an interesting group to study on three counts. First, other investigators have repeatedly found that earnings forecasts (and forecast revisions) have an important influence on stock prices (Philip Brown et al., 1985). Second, past work suggests that analysts are rather good at what they do. For example, analyst forecasts often outperform time-series models (see Robert Conroy and Robert Harris, 1987). Finally, the precision of analyst expectations represents a natural upper bound to the quality of the earnings forecasts of less sophisticated agents. After all, most investors do not have the time or the skill to produce their own predictions and, accordingly, they buy (rather than sell) earnings forecasts. Thus, for all of the above reasons, it is particularly interesting to see whether market professionals display any of the biases discovered in studies of nonexpert judgment.

[†] *Discussants*: Andrew Lo, MIT; Andrei Shleifer, University of Chicago; Fischer Black, Goldman Sachs.

[*] Cornell University and University of Wisconsin-Madison, and Cornell University, Ithaca, NY 14853, respectively. We thank Dale Berman from Lynch, Jones & Ryan and Bart Wear from First Wisconsin Asset Management for providing the data used in this study. We also gratefully acknowledge financial support from the Research Foundation at the Institute of Chartered Financial Analysts (De Bondt) and the Russell Sage Foundation (Thaler). Helpful comments have been provided by Josef Lakonishok, John Elliott, and Andrew Lo, though, of course, errors are our own responsibility.

We specifically test for a type of generalized overreaction, the tendency to make forecasts that are too extreme, given the predictive value of the information available to the forecaster. This tendency is well illustrated by an experiment conducted by Kahneman and Tversky. Subjects were asked to predict the future grade point average (GPA) for each of ten students on the basis of a percentile score of some predictor. Three predictor variables were used: percentile scores for GPA, for a test of mental concentration, and for a test of sense of humor. Obviously, a percentile measure of GPA is a much better predictor of actual GPA than is a measure of mental concentration which, in turn, is much more reliable than information on sense of humor. Therefore, subjects should give much more regressive forecasts in the latter two conditions, that is, the forecasts should be less variable. The results indicated that people were not nearly sensitive enough to this consideration. Subjects that were given a nearly useless predictor (the "sense of humor" condition) made predictions that were almost as extreme in variation as those given a nearly perfect predictor (the "percentile GPA" condition). This pattern leads to a systematic bias: forecasts that diverge the most from the mean will tend to be too extreme, implying that forecast errors are predictable.

This study asks whether security analysts display similar biases. Our focus is on forecasted changes in earnings per share (EPS) for one- and two-year time horizons. We study two questions. The first is whether forecast errors in EPS are systematically linked to forecasted changes. In particular, are the forecasts too extreme? Are most forecast revisions "up" ("down") if the analysts initially projected large declines (rises) in EPS? Clearly, under rationality, neither forecast errors nor forecast revisions should ever be predictable from forecasted changes. The second question is whether the bias in the forecasts gets stronger as uncertainty grows and less is (or objectively can be) known about the future.

Several of the regressions reported below are of the form $AC = \alpha + \beta FC$, where AC is the actual change and FC is the forecasted change. The null hypotheses of rational expectations is that $(\alpha, \beta) = (0, 1)$. The two alternative behavioral hypotheses sketched above are:

H1. Forecasted changes are too extreme, so actual changes are less (in absolute value) than predicted: $\beta < 1$.

H2. The estimated β for the two-year forecasts is less than the β for the one-year forecasts.

The next two sections describe the data and the empirical results. We find considerable support for the behavioral view. We then briefly discuss the sources of the systematic forecast error.

I. Data

The analysts' earnings forecasts are taken from the Institutional-Brokers-Estimate-System tapes (IBES) produced by Lynch, Jones & Ryan, member of the New York Stock Exchange. We study forecasts between 1976 and 1984. Lynch, Jones & Ryan contacts individual analysts on a regular basis and computes summary data such as means, medians, or standard deviations. The summary data that we analyze are sold to institutional investors. Updates are available each month but here we only work with the April and December predictions of EPS for the current as well as the subsequent year. The April forecasts are approximately one- and two-year forecasts since we only consider companies with a fiscal year ending in December. For these firms, actually realized earnings are typically announced sometime during the first few months of the following calendar year.

We match the earnings forecasts for each company with stock returns and accounting numbers. The returns are provided by the Center for Research on Security Prices (CRSP) at the University of Chicago. The accounting data are listed on the annual industrial (main and delisted) COMPUSTAT files, sold by Standard & Poor's. Since all data sources contain full historical records, no survivorship biases affect the sample selection. Care is taken to adjust for stock splits and stock dividends so that all current and past returns, earnings figures, and forecasts are expressed on a comparable basis. When necessary, forecasts of fully diluted EPS are converted to forecasts of primary EPS (excluding extraordinary items).

While some IBES data are available for approximately 2300 to 2800 companies each year, our annual sample contains many fewer observations. For example, for the one-year forecasts, the number varies between 461 and 785. This follows from the data selection criteria that we use. Companies only qualify if they have 1) records on IBES, CRSP, and COMPUSTAT; 2) returns on CRSP for three years prior to the forecast month; 3) EPS numbers on COMPUSTAT for ten years prior to the forecast month; 4) a December fiscal year; 5) the data needed to compute the variables in the regressions described below. Despite the stringent data requirements, our sample (in firm-years) is the largest we know of that has been used to study the rationality of earnings forecasts (compare Edwin Elton et al., 1984).

II. Methods and Results

Much of the regression analysis is based on three sets of variables: forecasted changes in EPS ($FC1$, $FC2$, and $FC12$), actual changes in EPS ($AC1$, $AC2$, and $AC12$), and forecast revisions ($FR1$, $FR2$, and $FR12$). The "consensus" one and two-year forecasts of earnings per share ($FEPS(t)$ and $FEPS\ (t+1)$) that we study are defined as the cross-sectional means or medians of analyst forecasts reported in April of year t ($t = 1976 \ldots 1984$). Forecasted changes are then computed as $FC1(t) = FEPS(t) - EPS(t-1)$, $FC2(t) = FEPS(t+1) - EPS(t-1)$, and $FC12(t) = FEPS(t+1) - FEPS(t)$, where $EPS(t)$ represents actually realized earnings per share. We compute actual earnings changes in a way that is similar to the forecasted changes. For example, $AC1(t) = EPS(t) - EPS(t-1)$. Eight-month forecast revisions ($FR1$) subtract the April forecast of $EPS(t)$ from the equivalent forecast in December. Twenty-month forecast revisions ($FR2$) are the difference between the December forecast in year $t+1$ and the April forecast of $EPS(t+1)$. Similarly, $FR12$ subtracts the April $FC12(t)$ from the equivalent $FC12(t)$ in December of year t.

TABLE 1—TESTS FOR THE RATIONALITY OF EARNINGS PER SHARE FORECASTS

Equation	Variables	Constant	Slope	Adj. R^2
1	$AC1$, $FC1$	−.094	.648	.217
		(−3.7)	(−21.7)	[5736]
2	$FR1$, $FC1$	−.120	−.181	.041
		(−6.7)	(−15.6)	[5736]
3	$AC2$, $FC2$	−.137	.459	.071
		(−2.3)	(−19.5)	[3539]
4	$FR2$, $FC2$	−.192	−.381	.074
		(−3.9)	(−16.8)	[3538]
5	$AC12$, $FC12$	.153	−.042	.000
		(2.4)	(−16.9)	[3520]
6	$FR12$, $FC12$	.348	−.439	.153
		(19.4)	(−25.3)	[3562]

Note: All variables are as defined in the text. The dependent variable is listed first. *T*-values appear in parentheses beneath the regression coefficients and test whether they differ from zero. However, for the slopes of equations 1, 3, and 5, the *t*-statistics test whether the coefficients differ from one. Note that the number of observations is given in brackets in the far right-hand column.

The regressions in Table 1 use mean consensus forecasts. All variables are normalized by the standard deviation of earnings per share between years $t-10$ and $t-2$.[1] Even though we also ran the regressions year by year, the results in Table 1 are based on the pooled samples. There are three main findings. Forecasts are too optimistic, too extreme, and even more extreme for two-year forecasts than for single-year predictions.

Equation 1 refers to the one-year forecasts. We regress the actual change in earnings on the April forecasted change. The intercept is significantly negative, indicating that the forecasts are too optimistic. This excessive optimism also appears in equation 3 for the two-year forecasts. The negative intercepts in equations 2 and 4 reveal that there is a general tendency for forecasts to be revised downwards between April and December.

---

[1] We also tried other normalization procedures, such as dividing by company assets per share at the end of year $t-1$, the stock price on the last trading day of year $t-5$, or the standard deviation of EPS between $t-5$ and $t-2$. Results are qualitatively the same for all methods.

The finding of unrealistic optimism seems consistent with the experimental research of Neil Weinstein (1980) and others who find such biases in the expectations of individuals in everyday life. However, we do not want to push this argument too far for two reasons. First, if we consider the nine individual year-by-year regressions, the intercepts are positive four times. Second, optimism bias also has a plausible agency interpretation. Many analysts work for brokerage houses that make money by encouraging trading. Since every customer is potentially interested in a buy recommendation, while only current stockholders (and a few willing to go short) are interested in sell recommendations, optimistic forecasts may be preferable. Indeed, it is well known that buy recommendations issued by brokerage houses greatly exceed sell recommendations.

All six regressions in Table 1 present evidence supporting the hypothesis that forecasts are too extreme. Ignoring the constant term in equation 1, actual EPS changes average only 65 percent of the forecasted one-year changes. For the two-year forecasts (equation 3), this statistic falls to 46 percent.[2] In the year-by-year regressions equivalent to equations 1 and 3, the slope coefficients are less than one every single time.

Note that equations 1 and 3 could be rewritten with the forecast errors ($AC1-FC1$ and $AC2-FC2$) on the left-hand side and with the forecasted changes as the regressors. The new slope coefficients then equal the *betas* in Table 1 *minus one*, while the t-statistics remain the same. The new slopes have a straightforward interpretation: The larger the forecasted changes, the larger is the forecast error in the opposite direction. The R^2s of these regressions are .076 and .097.

The previous findings all suggest that forecast revisions should also be predictable from forecasted changes, and indeed they are, as shown in equations 2 and 4. In these regressions, rationality implies that β should be equal to zero. In actuality, the slopes are significantly negative. By December, the average reversal of the one-year forecasts made in April equals 18 percent of the original predicted changes. For the two-year forecasts, the reversal amounts to 38 percent.

As expected, the results are stronger for the two-year and second year forecasts. The two-year results are clearly driven by the predicted changes for the second year (see equations 5 and 6). With the R^2 for equation 5 equal to zero, actual changes are simply unrelated to forecasted changes in EPS from year t to $t+1$. On average, any nonzero prediction, either positive or negative, is pure error. By December, the analysts have reversed their April forecasted changes for the second year by 44 percent.

In sum, the above results are consistent with generalized overreaction. However, a different interpretation is based on the problem of errors in variables. If our measure of forecasted changes in earnings contains error, then the slope coefficients are biased downward. In evaluating this argument, one should consider the most likely sources of error. One possibility is IBES data entry errors. Following Patricia O'Brien (1988), we removed any data points for which the predicted change in EPS or the forecast revision was greater than $10. The results in Table 1 reflect this error screen. In addition, we also recomputed regressions 1 and 3 using a smaller sample of firms for which the consensus forecast is based on the individual predictions of three analysts or more. For this subset, we then used the median forecasted earnings change as the regressor. The βs increased but were still significantly less than one.

A second potential source of error stems from the fact that the forecasts on the IBES tape may be stale. In fact, O Brien finds that the average forecast in the IBES sample is 34 days old. Stale forecasts are troublesome if the forecasters do not know the earnings for year $t-1$ when they make their predictions for year t. For example, a forecaster who thinks that year t earnings will remain unchanged from year $t-1$ will appear to be predicting a change in earnings if his estimate of $t-1$ earnings is wrong. We cannot completely rule out this interpretation of the results but we selected April as the month to study with an eye toward minimizing the problem. We chose the longest possible forecast horizon where we could still be reasonably confident that the forecasters would know the previous year's earnings. The April forecasts are issued in the third week in April so that, by O'Brien's estimate, the average forecast was made in mid-March. At this point, the analysts should either know the past year's earnings exactly or have a very good estimate. Thus it seems unlikely that such a large bias could be produced by errors of this type.

---

[2] The two-year regressions are open to criticism because the sampling interval (one year) is shorter than the forecast interval (two years) creating a nonindependence across data points. To remove this problem, we break the sample in two, and replicate equation 3 using forecasts just from every other year, so that the time periods are nonoverlapping. Results are comparable.

Another reason for confidence in the results reported here is that others have obtained similar results in previous studies of professional forecasters, both security analysts and economists. Using just the 1976–78 years of the IBES data, Elton et al. estimated regressions similar to ours, and obtained slope coefficients less than one in each year. In a study of exchange rate expectations, Kenneth Froot and Jeffrey Frankel (1989) also found evidence consistent with overreaction or, what they call, "excessive speculation." Forecast errors are regressed on forecasted changes in exchange rates. The slope coefficients which, under rationality, should equal zero, are always significantly different from zero. When an instrumental variables technique is used to correct for errors-in-variables, the results do not change. Finally, David Ahlers and Josef Lakonishok (1983) study economists' forecasts of ten macroeconomic variables, using the Livingston data set. In regressions similar to our equation 1, they find slope coefficients significantly less than one for each of ten variables being forecast. In other words, predicted changes were more volatile than actual changes, consistent with overreaction.[3]

We have documented generalized overreaction. However, an interesting question remains: What causes excessive optimism or pessimism in earnings forecasts? We considered several variables that might explain EPS forecast errors. Two variables that are of interest in light of our previous work include a measure of market valuation, MV/BV, the ratio of the market value of a company's equity to its book value (at the end of year $t-1$), and earnings trend (the growth rate of earnings over the years $t-6$ to $t-2$). Both variables were significantly related to forecast error in the expected direction, that is, excessive optimism for high MV/BV and high earnings growth firms, and excessive pessimism for firms low on these measures. Unfortunately, neither factor explained much of the variation in the forecast errors.

III. Conclusion

Formal economic models of financial markets typically assume that all agents in the economy are rational. While most economists recognize that, in fact, not everyone is fully rational, the existence of an irrational segment of the economy is often dismissed as irrelevant with the claim that there will be enough rational arbitrageurs to assure that rational equilibria will still obtain. Whatever the theoretical merits of this position (for a critique, see Bradford De Long et al., 1990; Thomas Russell and Thaler, 1985), an interesting empirical question is whether the presumed smart money segment actually can be identified. This paper investigates one possible source of rationality in financial markets, namely security analysts.

---

[3] As mentioned above, analysts may have incentives to make biased forecasts in order to stimulate trading by customers. Our discussant, Andrew Lo, suggested that these agency problems could produce an overreaction bias as well as an optimism bias. Whether or not this argument is plausible, the fact that the overreaction bias is observed for forecasters in domains in which the agency problem is not present suggests that the bias is produced by cognitive errors rater than faulty incentives.

ble source of rationality in financial markets, namely security analysts.

The conclusion we reach from our examination of analysts' forecasts is that they are decidedly human. The same pattern of overreaction found in the predictions of naive undergraduates is replicated in the predictions of stock market professionals. Forecasted changes are simply too extreme to be considered rational. The fact that the same pattern is observed in economists' forecasts of changes in exchange rates and macroeconomic variables adds force to the conclusion that generalized overreaction can pervade even the most professional of predictions.

The proper inference from this, we think, is to take seriously the behavioral explanations of anomalous financial market outcomes. When practitioners describe the recent October crashes as panics, produced by investor overreaction, perhaps they are right. After all, are not these practitioners the very same "smart money" that is supposed to keep markets rational?

REFERENCES

Ahlers, David and Lakonishok, Josef, "A Study of Economists' Consensus Forecasts," *Management Science*, October 1983, *29*, 1113–25.

Brown, Philip, Foster, George and Noreen, Eric, *Security Analyst Multi-Year Earnings Forecasts and the Capital Market*, Sarasota: American Accounting Association, 1985.

Conroy, Robert and Harris, Robert, "Consensus Forecasts of Corporate Earnings: Analysts' Forecasts and Time Series Methods," *Management Science*, June 1987, *33*, 725–38.

De Bondt, Werner F. M. and Thaler, Richard H., "Does the Stock Market Overreact?," *Journal of Finance*, July 1985, *40*, 793–805.

______ and ______, "Further Evidence on Investor Overreaction and Stock Market Seasonality," *Journal of Finance*, July 1987, *42*, 557–81.

De Long, Bradford et al., "Noise Trader Risk in Financial Markets," *Journal of Political Economy*, forthcoming 1990.

Elton, Edwin J., Gruber, Martin J. and Gultekin, Mustafa N., "Professional Expectations: Accuracy and Diagnosis of Errors," *Journal of Financial and Quantitative Analysis*, December 1984, *19*, 351–65.

French, Kenneth R. and Roll, Richard, "Stock Return Variances: The Arrival of Information and the Reaction of Traders," *Journal of Financial Economics*, September 1986, *17*, 5–26.

Froot, Kenneth A. and Frankel, Jeffrey A., "Forward Discount Bias: Is It an Exchange Risk Premium?," *Quarterly Journal of Economics*, February 1989, *104*, 139–61.

Kahneman, Daniel and Tversky, Amos, "On the Psychology of Prediction," *Psychological Review* 1973, *80*, 237–51.

O'Brien, Patricia C., "Analysts' Forecasts as Earnings Expectations," *Journal of Accounting and Economics*, January 1988, *10*, 53–83.

Russell, Thomas and Thaler, Richard H., "The Relevance of Quasi Rationality in Competitive Markets," *American Economic Review*, December 1985, *75*, 1071–82.

Shiller, Robert J., "Investor Behavior in the October 1987 Stock Market Crash: Survey Evidence," Working Paper, Cowles Foundation, Yale University, November 1987.

Weinstein, Neil, "Unrealistic Optimism about Future Life Events," *Journal of Personality and Social Psychology*, 1980, 806–20.

Market Volatility and Investor Behavior

By ROBERT J. SHILLER*

It appears that speculative asset prices tend to show excess volatility relative to simple present value efficient markets models, and that prices are partly forecastable as tending to returning to "mean," appropriately defined.[1] But what does this tell us about how speculative prices are determined? Are there "fads" in prices, and if so, how do they behave?

The question has arisen recently whether there is really room for fads in speculative prices. Whether or not speculative prices are too volatile, if stock prices are highly correlated with dividends we might conclude that the movements in stock prices are driven by fundamentals, not fads.

A question of longer standing is whether fads models that entail feedback from price change to price change are consistent with the observed approximate "random walk" price behavior, that is, rather low serial correlation of short-run price changes. We can also ask whether the feedback models are consistent with the observed relation of stock prices to dividends and earnings.

I. Is There Room for Fads in Speculative Prices?

If the detrended log real annual average Standard and Poor's Composite Stock Price Index 1871 to 1987 is regressed on detrended log real annual average real dividends, the R^2 is 0.637.[2] If a shorter 1951–87 sample is used, the R^2 is 0.819, even higher yet. This suggests that real stock prices are, for the most part, just reacting to real dividends (see my 1984 paper). Whether or not the reaction to dividends is appropriate from the standpoint of efficient markets theory, it would seem that we know that dividends are the ultimate source of most stock price movements.

Robert Barsky and Bradford De Long (1989) have argued that this correlation of prices and dividends is consistent with a model where investors are pricing stocks as present values of forecasted dividends, and forecasts are based on lagged real dividends, using an autoregressive forecasting model with two unit roots. Kenneth Froot and Maurice Obstfeld (1989) argued that this correlation is consistent with an "intrinsic bubble" model where price is infected with rational bubbles driven by dividend innovations.

Barsky and De Long make no argument that their forecasting model is the *correct* one for the dividend process we observe. Whether or not the model is correct, they find that the warranted price (forecasted present value) resembles actual real price a little better over the whole sample than does the real dividend series itself. The principal reason the R^2 in the regression noted above of price on dividends, 1871–1987, is not higher is that price is more responsive to dividends (and the price-dividend ratio higher) after 1950. Barsky and De Long's estimated forecasting model makes the pre-

---

*Cowles Foundation, Yale University, New Haven CT 06520.

[1] See my 1989 study. Those who criticize the various rejections of market efficiency have weakened the case against the efficient markets model, but the weight of the evidence remains in the direction of substantial excess volatility.

[2] In this paper, the nominal stock price index is the annual average S&P's Composite Stock Price Index from S&P's Statistical Service, *Security Price Index Record*, passim. The dividend and earnings series are the S&P series, per share, adjusted to index, 4-quarter total, and extended back from 1926 to 1871 using Cowles data, appearing in Table 26.1, Series 2 and 3, in my 1989 study. These series are deflated by dividing by an annual average producer price index (Table 26.2, Series 7, in my 1989 study). Detrending is done by regressing the log real series on a linear time trend and a constant, and taking the residual.

dicted growth rate of real dividends a long distributed lag on past growth rates. Hence, the warranted price-dividend ratio is higher near the end of the sample when real dividends were higher. Froot and Obstfeld's model has a rational bubble that is an upwardly concave function of dividends themselves; this allows their warranted real price to capture the post-1950 behavior of real stock prices by making it more responsive to real dividends when dividends were higher, near the end of the sample. But over most of the sample, neither of these warranted real prices corresponds much more closely to real prices than do real dividends themselves.

I offer here two reasons why there is substantially more room for fashions and fads in stock prices than is suggested by the correlation between stock prices and dividends.

First, the correlation between price and dividend described above is substantially a low-frequency one: dividends show long-run swings, and price tracks these somewhat. But at higher frequencies, the correlation between price and dividend is much lower. If the annual change in log real stock price is regressed on the contemporaneous change in log real dividends 1871 to 1987, the R^2 is only 0.301. For the subsample 1951–87, the R^2 is only 0.231. Thus, most of the year-to-year change in log real price is not explained by changes in log real dividends. Adding a few more lagged changes in log real dividends to the regression does not substantially affect the R^2, so no forecasting model for the present value of dividends that is linear in these log real dividends will fit price changes better.

The second point is that dividends may themselves incorporate some of the fashions and fads that infect stock prices, so that even the component of price that is correlated with dividends may be influenced by fashions and fads. Managers set dividends; dividends are under their discretion. If managers become optimistic at the same time that the public does, they may increase dividends at the same time that fashions cause prices to increase. Some evidence on the potential importance of this point can be had by regressing detrended log real price on the contemporaneous detrended log real earnings. Earnings are less likely to be affected by changes in optimism or pessimism than are dividends, since the former are restricted by standard accounting practices. Using the sample 1871–1987, if detrended log real price is regressed on detrended log real earnings, the R^2 is 0.318, With data 1951–87, the R^2 in this regression is 0.341. When the annual change in log real price is regressed on the annual change in log real earnings 1871–1987, the R^2 is only 0.170; with the sample 1951–87, the R^2 is only 0.145. Again, adding lagged earnings changes to the latter regression does not substantially improve the R^2. One might say that hardly any of the year-to-year change in price can be explained in terms of the exogenous earnings data.

Someone might still argue that the earnings series is a noisy one, and that the dividend series, rather than being influenced by fashions and fads, is itself a better measure of the income of the firm than are earnings themselves. Terry Marsh and Robert Merton (1986) have presented a rational expectations model of dividend-setting behavior that has managers responding to "permanent earnings," their long-run forecast of future earnings.

If Marsh and Merton are right, that dividends are set by managers in anticipation of future earnings, then any changes of dividends relative to earnings reflects information about future earnings changes. If managers have any information about future earnings other than that contained just in lagged earnings, this would mean that changes in dividends should Granger-cause changes in earnings.

Granger causality tests do not reject the hypothesis that dividends do not cause earnings. A Granger causality test of no causality from the annual change in log real dividends to the annual change in log real earnings was conducted using the data 1871–1987 (the longest sample possible) with 10 lags. The test yielded an F-statistic of 0.886 with 10, 85 degrees of freedom, that, with p value of 0.549, is not significant. Thus, the movements in log real dividends that are uncorrelated with current and lagged log real earnings do not appear to help forecast real earnings changes. The theory that dividends are an indicator of future earnings is unsupported; there is no evidence here against the notion that the correlation of price with dividend is due to the common effect of fads on both.[3]

II. Can Feedback Models Fit the Data?

The standard feedback theory for speculative bubbles is that when market prices as a whole rise substantially, this creates many success stories of investors, and these stories entice potential investors, who naively imagine that the same success will come to them if they invest too. Substantial decreases in market prices create many stories of investors' failures, and these discouraging stories drive investors away from the market. Probably such feedback stories seem unattractive to economists, since they posit the most extreme form of naivete.[4] But causal impressions are suggestive that such naivete *does* play a role in market behavior. The periods of the bull markets of the 1920s and 1950s were times when many observers thought that investors were being enticed into the stock market this way, and the periods after the crashes of 1929 or 1987 were periods when many observers also thought that investors were being scared away from the market in this way.

These popular impressions of the attitudes of investors are hardly convincing evidence, and so it is natural to seek some statistical evidence from the data that bears on their validity. Now, it is common to assume that such feedback models would imply high positive serial correlation of short-run price changes. If so, the models would seem to fly in the face of statistical evidence, since in fact serial correlation of short-run price changes is known to be low. This argument has been raised often against feedback models, but the argument has not been examined critically.

Such feedback models do not tend to create high serial correlation of short-run price changes since the feedback may operate over long time intervals (i.e., with a distributed lag). Let us consider a distributed feedback model that illustrates how feedback might work. The model has log real price p_t determined by its own lagged changes and a forcing variable π_t:

$$p_t = \pi_t + c\int_{-\infty}^{t} e^{-\gamma(t-\tau)}\,dp_\tau, \qquad 0<\gamma, \quad 0<c<1 \tag{1}$$

The integral in this expression is a (continuous time) distributed lag on past rates of price change dp_τ. The forcing variable π_t is the valuation people would place on a stock if there were no feedback through price. Since c is positive, the model represents that past price increases encourage investment and thus boost price, past decreases discourage investment and thus diminish price. The parameter c must be less than one if price is not to explode; presumably investors are not so irrational as to let this happen. Since γ is greater than zero, the more distant past is given less weight (reflecting gradual memory loss) in determining p_t than is the more recent past. We shall, however, suppose that γ is small enough that price changes for many years past still matter today.

Equation (1) can be solved to give p_t in terms of π_t:

$$\begin{aligned} p_t &= \pi_t + \frac{c}{1-c}(\pi_t - \bar{\pi}_t) \\ &= \frac{1}{1-c}\pi_t - \frac{c}{1-c}\bar{\pi}_t \end{aligned} \tag{2}$$

where

$$\begin{aligned} \bar{\pi}_t &= \lambda\int_{-\infty}^{t} e^{-\lambda(t-\tau)}\pi_\tau\,d\tau \\ \lambda &= \gamma/(1-c). \end{aligned} \tag{3}$$

Here, $\bar{\pi}_t$ is a (long) weighted average of lagged π. This model implies that price p_t is cointegrated with π_t, but amplifies its departures from a weighted average of its own lagged values. Log price p_t will be higher than π_t if π has been increasing in recent years, lower if decreasing.

---

[3] We can also look at time-series on farm rents (i.e., dividends in our model) and farm prices. Farm rents would seem less likely than corporate dividends to be afflicted by fads in prices. Here the evidence that price is driven by rents appears mixed. Barry Falk (1990), with Iowa data, finds a substantial correspondence between price and rent; Mark Carey (1989) with Illinois data does not. (Both reject the efficient markets model.)

[4] De Long et al. (1989) show that the presence of "smart money" may not serve to reduce the effect of feedback traders on price, but to enhance it.

Over short intervals of time, $\bar{\pi}_t$ is virtually a constant, and so p_t has essentially the same short-term stochastic properties as π_t, but amplified. Suppose π_t is a Wiener process (a continuous time random walk), as would occur if public opinion about stocks were, in the absence of feedback, shocked randomly by news events. Then although feedback amplifies the movements in price, price is still approximately a random walk. (John Pound and I, 1989, advanced a similar argument why interpersonal contagion models of investor interest might imply that π_t is approximately a random walk.)

This story is oversimplified in that it has no response of price to dividends. Consider now a variation on (1). Let us replace π_t with log real dividends plus π_t: $d_t + \pi_t$. If dividends are not influenced by fads and if fads are truly capricious, then d_t might be expected to be uncorrelated, at least roughly, with dividends at all lags. Under this assumption, suppose we regress log real price on current and lagged log real dividends. The model (1) would imply that the coefficient of current dividend should be $1/(1-c)$ and greater than 1, the coefficients of lagged dividends are negative. When we run a regression of log real price on current log real dividend and a 30-year moving average of lagged log real dividends using our data 1901–87, the estimated coefficient of the current log real dividend is 1.609, and the coefficient of the 30-year average of lagged log real dividends is -0.257. If data 1950–87 are used, the coefficient of the current log real dividend is 2.203 and the coefficient of the average of lagged log real dividends is -0.666. (The apparent "overreaction" to dividends is also reflected in John Campbell's 1990 observation that news about future dividends is negatively correlated with news about future returns.)

It was, however, argued above that fads in stock pricing might be communicated to dividends through the dividend-setting behavior of managers. A positive correlation of fads with dividends might then mean that the estimated coefficient of dividend is biased upward, so that people are not in fact reacting to dividends despite the correlation of prices and dividends. We might consider then replacing $d_t + \pi_t$ with e_t (log real earnings) $+ \pi_t$. When we regress p_t on e_t and a 30-year moving average of lagged e_t (1901–87), the coefficient of current log real earnings is less than 1, at 0.822, and the coefficient of the 30-year moving average is 0.132. Prices do not appear to amplify departures of earnings from the 30-year moving average of earnings as they do for dividends; but then certainly earnings are a noisier series than are dividends. On the other hand, with data 1950–87, the coefficient of the current log real earnings is 1.429 and the coefficient of the 30-year moving average is -0.505, confirming for this sample the pattern of coefficients suggested by the simple feedback model.

III. Interpretation

The observed volatility of speculative prices and the pattern of feedback of price to dividends or earnings is suggestive of simple feedback models of investor behavior. But these simple feedback models are not at this time well established nor are they satisfactory theoretical constructs on which to base our understanding of financial markets.

Further research on a couple of fronts would help us to advance such models. Research on the relation of fundamentals, such as demographics, to prices would appear to be worthwhile. Zhi-Wu Chen (1989) finds that a baby-boom-related demographic variable substantially forecasts long-horizon returns in the U.S. stock market, 1948–87. The average age of the population over age 20 shows a hump shape (rising in the 1950s, peaking in the late 1960s, falling until the early 1980s) that is rather similar to the hump in real stock prices (also rising on average between the 1950s and the peak in the late 1960s and then falling on average until the early 1980s). His evidence, then, suggests that fads might have less of a role in price movements than the above analysis of prices, dividends, and earnings would allow. N. Gregory Mankiw and David Weil (1989) found much of the behavior of a real housing price measure, the residential investment deflator divided by the GNP deflator in the United States in the postwar period, could be explained in terms of a different baby-boom variable which is essentially detrended adult population. Their real housing price shows an *inverted* hump in the postwar period, roughly a mirror image of that in the stock market, real housing prices falling in the 1950s, bottoming in the late 1960s, and rising through the 1970s. Their different baby-boom variable roughly captures this. Further studies like these covering longer time periods and more markets might reveal more convincingly the importance of such fundamentals for speculative prices.

Further research on the ever-changing pool of popular theories and models is another useful direction to take. Investors approach trading in speculative markets in an attentive and intellectual way. They have *models* of market behavior; they respond to news and information. But people are not pursuing, or capable of pursuing, a proper decision-theoretic analysis of the evidence for all these competing models. Thus, many competing popular models are outstanding at any time, some people believing one and others believing others. Price increases do not automatically feed back into investor demand as the above feedback model would suggest. Rather, we should regard any tendency of feedback of price changes into price as reflective, instead, of a tendency for bias in the popular models toward models that encourage investors to react positively to price increases and negatively to decreases.

Rather than look only at aggregated data, and the effects of these models as a group, the information on the popular models themselves will allow more adequate theorizing about human behavior in speculative markets.[5]

---

[5] I with several coauthors have begun collecting data on popular models in connection with the stock market and market for housing (see my 1989 study).

REFERENCES

Barsky, Robert and De Long, J. Bradford, "Why Have Stock Prices Fluctuated?," reproduced, University of Michigan, July, 1989.

Campbell, John Y., "Measuring the Persistence of Expected Returns," *American Economic Review Proceedings*, May 1990, *80*, 43–47.

Carey, Mark, "Feeding the Optimists: The Federal Land Banks, Land Market Efficiency, and the Farm Credit Crisis," reproduced, UC-Berkeley, 1989.

Chen, Zhi-Wu, "Population Aging and Expected Market Risk Premiums," reproduced, Yale University, October 1989.

De Long, J. Bradford et al., "Positive Feedback Investment Strategies and Destabilizing Rational Speculation," reproduced, Harvard University, February 1989.

Falk, Barry, "Formally Testing the Present Value Model of Farmland Prices," *American Journal of Agricultural Economics*, forthcoming, 1990.

Froot, Kenneth A, and Obstfeld, Maurice, "Intrinsic Bubbles: The Case of Stock Prices," reproduced, MIT, 1989.

Mankiw, N. Gregory and Weil, David, "The Baby Boom, the Baby Bust, and the Housing Market," *Regional Science and Urban Economics*, May 1989, *19*, 235–58.

Marsh, Terry A. and Merton, Robert C., "Dividend Variability and Variance Bounds Tests for the Rationality of Stock Market Prices," *American Economic Review*, June 1986, *76*, 483–98.

Shiller, Robert J., *Market Volatility*, Cambridge: MIT Press, 1989.

______, "Stock Prices and Social Dynamics," *Brookings Papers on Economic Activity*, 2:1984, 457–98.

______ **and Pound, John,** "Survey Evidence on Diffusion of Interest and Information among Investors," *Journal of Economic Behavior and Organization*, Fall 1988, *12*, 47–66.

Speculative Dynamics and the Role of Feedback Traders

By DAVID M. CUTLER, JAMES M. POTERBA, AND LAWRENCE H. SUMMERS*

The 1987 stock market crash demonstrated more convincingly than any econometric test ever could that not all movements in asset prices can be accounted for by news about fundamental values. The efficient markets hypothesis was probably the right place for serious research on asset valuation to begin, but it may be the wrong place for it to end. In this paper, we review some of our research directed at providing an alternative framework for thinking about fluctuations in speculative prices.

As proponents of the efficient markets hypothesis stress, repeated analysis of the single time-series on U.S. stock returns is bound to turn up patterns sooner or later. Our research has therefore sought to determine whether there are regularities that appear not just in U.S. equity returns, but also in returns in other countries' stock markets, and in other assets. Given that risk factors are likely to operate differently in different markets, finding common patterns across markets suggests the need for consideration of the speculative process itself.

After summarizing our earlier research documenting the characteristic speculative dynamics of many asset markets, we go on to suggest a framework for understanding them. Our model incorporates "feedback traders," traders whose demand is based on the history of past returns rather than the expectation of future fundamentals. We use this framework to describe ways in which the characteristic return patterns might be generated, and also to address the long-standing question of whether profitable speculation stabilizes asset markets.

I. Characteristic Speculative Dynamics

Table 1, which is drawn from our earlier paper (where the results are described in much greater detail) provides summary evidence on three empirical regularities in the markets for stocks, bonds, foreign exchange, and various real assets. First, excess returns display positive autocorrelation at relatively short horizons. Both the 1-month return autocorrelation (col. 1) and the average of the first 12 monthly autocorrelations (col. 2) are positive and statistically significant. The average 1-month serial correlation coefficient for the 13 equity markets we consider exceeds .10, and bond markets exhibit even greater autocorrelation.

Second, there is a weak tendency for returns to be negatively autocorrelated at durations of several years. The average autocorrelations at 13–24 months (col. 3) are negative for stocks, bonds, and foreign exchange, although the latter finding is not statistically significant.

Many technical trading systems are designed to take advantage of exactly the sort of serial correlation patterns suggested here. It can be shown, for example, that procedures that involve using the crossing of two moving averages as a trading signal are optimal if autocorrelations are at first positive and then negative.

Third, in most cases, returns over periods of several years can be predicted on the basis of crude proxies for the deviation of asset prices from fundamental value. For each market, we defined a proxy "fundamental value": a constant multiple of dividends in

---

*Cutler and Poterba: MIT, Cambridge, MA 02139; Summers: Harvard University, Cambridge, MA 02138. This paper draws on our earlier joint work (1990). Our discussion of destabilizing speculation is largely derivative of J. Bradford De Long et al. (1990). We are grateful to Fischer Black and John Campbell for helpful comments, and to the NSF and Alfred P. Sloan Foundation for research support. This research is part of the NBER Research Program in Financial Markets and Monetary Economics and Economic Fluctuations.

TABLE 1—PROPERTIES OF ASSET EXCESS RETURNS

Asset	(1)	(2)	(3)	(4)
Equities (1960–88,	.101	.021	−.026	41.20
13 markets)	(.026)	(.006)	(.008)	[27.48]
Bonds (1960–88,	.238	.064	−.013	−10.21
13 markets)	(.041)	(.011)	(.005)	[17.08]
Ex. Rates (1974–88,	.067	.033	−.010	31.92
10 markets)	(.037)	(.012)	(.012)	[18.43]
Gold (1974–88)	.020	.051	.017	133.19
	(.075)	(.022)	(.023)	[20.98]
Houses (1970–86,	–	.206	.083	–
4 cities)		(.032)	(.033)	
Collectibles (1968–88,	–	.365	.011	–
7 markets)		(.160)	(.153)	

Notes: Cols. 1–3 report autocorrelation of excess returns: Col. 1 presents the 1-month autocorrelation; col. 2 the 1–12-month average autocorrelation; col. 3 the 13–24-month average autocorrelation. Col. 4 denotes the Percent Reversion to Fundamentals within 4 Years.

The excess returns on all assets are measured as nominal returns less the short-term interest rate. Returns are monthly for all assets except houses (quarterly) and collectibles (annual). For all assets except gold, the standard errors in parentheses correspond to the standard error of the average autocorrelation across markets. The regression coefficient in col. 4 is the result of regressing the 48-month return on the logarithm of the ratio of fundamental value to the current asset price as the explanatory variable. The fundamentals are defined for each asset in the text. The values in brackets are the probability of observing regression coefficients at least as positive as the reported value. These *p*-values are based on Monte Carlo simulations described in detail in our earlier paper.

the case of stocks, the reciprocal of the short-term interest rate for bonds, and a constant real exchange rate and real gold price. Column 4 reports regression coefficients from equations relating subsequent 48-month excess returns to the logarithm of the current fundamental-price ratio. In the markets for equities and gold, and to a lesser extent foreign exchange, these measures have substantial forecast power for returns. The result for equities suggests that 40 percent of the deviation between price and our fundamental value measure is eliminated within 48 months. Similar evidence for house prices is presented by Karl Case and Robert Shiller (1989), who show that the real capital gain on houses can be forecast using lagged values of the rental-to-price ratio. They also show that over horizons longer than those in Table 1, the real capital gains on houses exhibit negative serial correlation.

Changing risk factors have thus far been unable to explain these characteristic patterns of asset returns. For several reasons, we suspect that theories focusing on the dynamics of speculation will be more successful. First, the pattern of correlations is similar in markets where risk might be expected to operate very differently, for example the bond and stock markets. Indeed, in the foreign exchange market, risk affects both currencies and thus has no predictable effect on exchange rate levels.

Second, qualitative discussions of major movements in speculative prices (for example, Charles Kindleberger, 1978, or Elliot Montrell and Wade Badger, 1974) focus on the interaction between traders who extrapolate past price increases and traders whose expectations are formed on the basis of fundamentals. We suspect that such accounts may also explain more recent movements in asset prices. In the summer of 1987, for example, stock prices were near record highs relative to dividends or earnings. Although this could be attributed to investor perceptions that equities were safer than they had been in the past, a more plausible account is that investor demand for equities was fueled by expected large capital gains from a continuing bull market.

Third, Poterba and Summers (1988) show that for specifications of the equity risk process that are consistent with empirical findings on volatility, increases in risk that raise future required returns should reduce current returns, thus leading to *negative* autocorrelation at high frequencies. Finally, John Campbell and Shiller (1988) show that there is little evidence that fluctuations in the price-dividend ratio forecast increases in real interest rates or other measurable aspects of risk.

II. Modeling Speculative Dynamics

This section develops a model of asset price dynamics when investors follow heterogeneous trading strategies. The importance of investor differences is clear from the substantial volume in modern securities markets, far more than would be expected if all investors held market portfolios and traded only to rebalance or finance consumption outlays. On the New York Stock Exchange, for example, almost 75 percent of the shares trade hands each year (New York Stock Exchange, 1988), and it is estimated that almost $400 billion of foreign currency is traded each day.

We consider a futures market, where there is a well-defined fundamental equal to the terminal value, but where there are no dividend payments. We also assume that the asset is in zero net supply.

We postulate three types of traders. The first group invest on the basis of rational forecasts of future returns, holding a higher fraction of the speculative asset when expected returns are high:

$$s_{1,t} = \gamma(E_t R_{t+1} - \rho); \quad \gamma > 0, \tag{1}$$

where R_t is the *ex post* return in period t, E_t is the expectation operator using information available as of time t, and ρ is the required return on the risky asset. For sufficiently large γ, this model reduces to the traditional constant required return model of asset pricing.

The second class of investors, fundamentals traders, base expected returns on prices relative to perceived fundamentals. When prices are high relative to perceived fundamentals, their demand is low. If the logarithm of the price and (true) fundamental value are respectively p_t and f_t, demand is

$$s_{2,t} = \beta(p_t - \alpha(L)f_t); \quad \beta < 0; \quad \alpha(1) = 1. \tag{2}$$

Such behavior is implied by investment strategies such as those based on "dividend discount models." We assume that the logarithm of the fundamental evolves as a random walk: $f_t = f_{t-1} + \varepsilon_t$. We allow for the possibility that perceived fundamentals reflect true fundamentals with a lag (i.e., that $\alpha(L)$ does not equal unity). If some traders have quicker access to information than others, in any period only some of the traders will know the current state of fundamentals.

Finally, feedback traders base demand on past returns:

$$s_{3,t} = \delta(L)(R_t - \rho) \tag{3}$$

where $\delta(L)$ is an arbitrary lag polynomial. Positive feedback trading, buying after price increases, could result from the use of stop loss orders, from portfolio insurance, from a positive wealth elasticity of demand for risky assets, or from margin call-induced selling after periods of low returns. It could also result from technical analysis models designed to catch incipient trends. Negative feedback trading, buying after price declines, could result from "profit taking" as markets rise, or from investment rules that target a constant share of wealth in different assets.

Asset market equilibrium requires

$$s_{1,t} + s_{2,t} + s_{3,t} = 0. \tag{4}$$

Assuming a constant required return (ρ) of zero, this yields a rational expectations difference equation for the asset price:

$$E_t(p_{t+1} - p_t) = -\beta(p_t - \alpha(L)f_t)/\gamma - \delta(L)(p_t - p_{t-1})/\gamma. \tag{5}$$

Solving this equation gives the asset price as a function of past prices, expected future fundamentals, and past fundamentals. The pricing function also displays the property that fundamental innovations (ε_t) are ultimately fully reflected in prices.

III. Explaining the Stylized Facts

This model can generate positive serial correlation in returns in any of three ways. First, if fundamentals traders learn about true fundamentals with a lag ($\alpha(L) \neq 1$), then fundamentals perceived by these traders will differ from those perceived (correctly) by the rational traders. With no feedback traders ($\delta(L) = 0$), this implies that following positive news, the rational traders will drive the price above perceived fundamentals of the fundamentals traders ($\alpha(L)f_t$) but below the true fundamental (f_t). The expected capital gain to the rational traders when news is ultimately incorporated in prices is just enough to induce them to hold a long position opposite the fundamentals traders. As the information is incorporated, the fundamentals traders will purchase assets held by the rational traders. Returns will therefore be positively serially correlated for as many periods as it takes for the information to be incorporated in demand.

Negative feedback traders ($\delta(L)<0$) are a second potential source of positive autocorrelation. Consider the market without fundamentals traders ($\beta=0$) and with 1-period negative feedback traders ($\delta_1<0$). Positive returns associated with favorable shocks to fundamentals reduce asset demand from negative feedback investors. Rational investors must take offsetting long positions, so expected returns to these investors must rise. On average, subsequent returns will therefore be higher, and returns will be positively serially correlated. The higher returns will show up as capital gains on the asset so that, as with the earlier case, the initial price reaction to the news will be incomplete. Negative feedback trading by central banks "leaning into the wind" to delay the incorporation of news into exchange rates has been advanced as a possible explanation for positive autocorrelation in currency returns.

Finally, positive autocorrelation can result from the presence of feedback traders who respond to returns in several previous periods. If excess returns in one period affect feedback trader demand in many subsequent periods, feedback traders will persistently demand long or short positions. Required returns for rational investors will therefore be above or below average for several periods, and this pattern will be reflected in positively correlated *ex post* returns. The precise autocorrelation properties generated by this model depend both on the nature of the feedback demand and on the speed with which fundamentals traders incorporate news about fundamentals into demand. Even without fundamentals traders, however, slowly adjusting positive feedback traders can induce positive autocorrelation of returns.

This third scenario, feedback traders with long memories, can generate negative autocorrelation at longer horizons as well as short-run positive autocorrelation. With enough positive feedback demand, prices will overreact to fundamental news. In the long run, however, prices must change by only the amount of the fundamental shock. This implies that returns must be negatively serially correlated over some horizons.

IV. Can Profitable Speculation Lead to Instability?

Models with heterogeneous traders can be used to study a variety of issues concerning the performance of asset markets. We illustrate this by examining the effects of speculation on price stability. The traditional view (presented, for example, by Milton Friedman, 1953), holds that profitable speculation (buying when prices are low and selling when they are high) will offset other market shocks and thereby stabilize prices. De Long et al., however, show that this view may be incorrect when some market participants engage in feedback trading. They present a stylized model in which profitable speculation can *raise* the variance of returns relative to the variance of shocks to fundamental values.

The framework developed above can be used to illustrate this point. Table 2 presents the results of numerical solutions of the autocorrelations, the variance of returns, and the variance of the fundamental-price deviation for the following special case of the model described above:

$$(6)\quad s_{1,t}=E_t p_{t+1}-p_t$$

$$(7)\quad s_{2,t}=-.25(p_t-.75f_t-.25f_{t-1})$$

$$(8)\quad s_{3,t}=.05(p_t-p_{t-1})+.05(p_{t-1}-p_{t-2})+.05(p_{t-2}-p_{t-3}).$$

As Table 2 shows, returns exhibit positive first-order autocorrelation, but are negatively serially correlated at two and three lags.

For this case, a speculator following a positive-feedback investment rule over short

TABLE 2—AN EXAMPLE OF DESTABILIZING SPECULATION

Summary Statistic	Base Case	Properties After Increasing γ	δ_1	α
Autocorrelations				
Lag 1	.0220	.0196	.0137	.0200
Lag 2	−.0347	−.0324	−.0343	−.0346
Lag 3	−.0405	−.0374	−.0401	−.0405
Variance of Returns	1.106	1.100	1.124	1.110
Variance of Price Around Fundamental Value	.0092	.0081	.0100	.0094

Note: This table reports numerical solutions of the 3-equation model in (6)–(8). The last three columns increase the indicated parameters by .10 (for γ) and .01 (for δ_1 and α). The variance of returns and the variance of the deviation between price and fundamental value are scaled by the variance of fundamental innovations.

horizons would earn profits. Table 2 shows that while an increase in this type of speculation (an increase in δ_1) reduces the serial correlation in returns, it *raises* their variability. In this example, prices initially undershoot changes in fundamentals. An increase in the importance of short-horizon feedback trading brings prices closer to fundamentals, but also increases feedback demand in subsequent periods. This increases the variance of returns as well as the variance of prices around fundamentals. Profitable speculation is therefore destabilizing.

Table 2 also reports comparative static results for two other parameter changes. An increase in γ, that raises the responsiveness of rational traders to changes in expected returns, moves all autocorrelations toward zero and reduces the variance of returns and the variance of prices around fundamentals. Raising α, the speed with which fundamentals traders incorporate information into prices, also reduces the autocorrelations but destabilizes prices for reasons similar to those above.

V. Conclusion

Our analysis of how feedback traders affect asset returns assumes that investors do not learn from past experience. A more realistic model would allow trading rules to change in response to factors such as the recent success of different portfolio strategies or, as in Robert Barsky and DeLong (1989), new information about the stochastic process of dividends or prices. For example, if investors inferred from the prewar experience that stock prices were negatively serially correlated at long horizons, and so rushed to purchase at troughs and to sell at peaks, they would reduce this serial correlation. Such adaptive trading rules would generate time-varying properties for asset returns and may provide a partial explanation for the finding, emphasized by Myung Kim et al. (1989), that the negative serial correlation in long-horizon U.S. stock returns is more pronounced prior to World War II than in the subsequent period.

REFERENCES

Barsky, Robert and De Long, J. Bradford, "Why Have Stock Prices Fluctuated?," mimeo., Harvard University, 1989.

Campbell, John Y. and Shiller, Robert, "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, Fall 1988, *1*, 195–228.

Case, Karl E. and Shiller, Robert J., "Forecasting Prices and Excess Returns in the Housing Market," mimeo., Yale University, 1989.

Cutler, David M., Poterba, James M. and Summers, Lawrence H., "Speculative Dynamics," NBER Working Paper No. 3242, January 1990.

De Long, J. Bradford et al., "Positive Feedback Investment Strategies and Destabilizing Speculation," *Journal of Finance*, forthcoming 1990.

Friedman, Milton, "The Case for Flexible Exchange Rates," in *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Kim, Myung, Nelson, Charles and Startz, Richard, "Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence," mimeo., University of Wash-ington, 1989.

Kindleberger, Charles P., *Manias, Panics, and*

Crashes, New York: Basic Books, 1978.

Montrell, Elliot W. and Badger, Wade W., *Introduction to Quantitative Aspects of Social Phenomena*, New York: Gordon & Breach, 1974.

Poterba, James M. and Summers, Lawrence H., "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics* September 1988, *22*, 27–60.

New York Stock Exchange, *Fact Book 1988*, New York: NYSE, 1988.

THE ECONOMICS AND POLITICS OF BUDGET CONTROL[†]

Budgetary Pressures in the EEC: A Fiscal Federalism Perspective

By DAVID E. WILDASIN*

As is widely known, the EEC has gone through some troubling budgetary difficulties in recent years. Disputes have revolved around the growth of agricultural subsidies and about the contributions to be made to the EEC by various member countries; in particular, the U.K. These problems have been brought under control for the short run. However, the implementation of the Single European Act could be greatly complicated, and might falter, if policymakers again became intensely involved in acrimonious budgetary disputes. Maintaining budgetary control for the next several years, and for that matter in the longer term as well, is thus a matter of considerable tactical importance within the context of the overall progress of European economic affairs. In view of the fact that the threat of acute budgetary crisis has receded somewhat, and in view of the prospect of a significant step forward toward more complete economic integration associated with the 1992 initiative, it seems appropriate to give some thought to the budgetary problems that the EEC is likely to confront in the medium term.

At a time of enormous political change in Eastern Europe, it is exceptionally hazardous to attempt projections about economic affairs in the EEC. It is easy to visualize scenarios in which significant amounts of resources, whether from the EEC itself or from individual member states, are directed toward promoting economic and other reform in the East, forcing new choices with respect to the commitment of resources by member countries to the EEC. However, to limit the scope of this paper, attention is restricted to developments within the EEC itself.

The major trends in the EEC budget are easily delineated, since only a few main initiatives account for the bulk of EEC expenditures. First, agricultural subsidies through the European Agricultural Guidance and Guarantee Fund (EAGGF) have entailed large expenditures. Through the 1980s, these outlays accounted for about two-thirds of EEC expenditures. Another major category of EEC expenditures are for the so-called "structural funds," in particular the European Regional Development Fund (ERDF) and the European Social Fund (ESF). The ERDF funds economic development projects in specially targeted low-income regions. The ESF funds programs that promote the training and employment of workers, especially youth and long-term unemployed persons. These and related expenditures have accounted for about 10–15 percent of expenditures during the 1980s. Altogether, then, agricultural and structural expenditures account for about 85 percent of the total budget. The U.K. has repeatedly protested against what it regards as excessive contributions to the EEC relative to the return that it gets in the form of structural fund outlays, and it has garnered a partial rebate of its contributions to the EEC, equal to two-thirds of the difference between the U.K. VAT contribution and EEC expenditures allocable to the U.K. These rebates account for roughly 5 percent of the EEC budget.

As a result of decisions taken in 1988, the outline of the EEC budget for the next several years is relatively clear. The special arrangements regarding U.K. contributions are to remain essentially intact. The growth of EAGGF expenditures is to be limited while ESF and ERDF expenditures are to be doubled (in real terms) by 1992. Expenditures on agricultural guarantees should fall to about 55 percent of the total budget in 1992, while expenditures in the structural funds will rise to about 25 percent. A notable feature of the evolving EEC budget is the extent to which it has focused resources on redistributional policies. For example, the expenditures in the structural funds are explicitly geared toward assistance in low-income areas of the Community.

[†] *Discussants*: Paul Courant, University of Michigan; Rudolph Penner, The Urban Institute; John Yinger, Syracuse University.

*Department of Economics, Indiana University, Bloomington, IN 47405. I am grateful to the Department of West European Studies at Indiana University for support of my research. A longer version of this paper is available on request.

For the purposes of budgetary analysis, the EEC is usefully viewed as an emerging federation. The EEC institutions can be seen as a central governmental agency that obtains resources from lower-level governments (the member states) and expends these resources through a system of grants. As we have seen, most of these grants support programs with redistributional objectives. Familiar principles of fiscal federalism assert that redistribution in a federal system creates externalities that warrant intervention by a higher-level government through appropriate grant policies or a complete shift in the responsibility for redistributive activities from lower- to higher-level governments. (See Richard Musgrave, 1969, and Wallace Oates, 1968, 1972.) Thus, it is not surprising to see the EEC budget evolving along the lines that we have observed. However, these redistributive activities deserve more thorough consideration. This paper outlines two simple models which can be used to analyze EEC expenditures, focusing on the economic and political linkages within and among the member states. I argue that these linkages can play a critical role in determining the impact of EEC grants.

I. Two Models of Redistribution with Spillovers

A. *A Model with Altruistic Spillovers*

In all European countries, governments undertake substantial economic interventions that favor particular regions, industries, or occupational groups. These programs constitute a bundle or package of assistance that implies some set of net transfers from the rest of the country to particular industries or regions. Think of these policies as the equilibrium of some political process. EEC grants perturb this equilibrium. What is the nature of the response?

Let us focus on a particular region in a country that will be the target of assistance from the ERDF for, say, a water project. Imagine an initial equilibrium in which individuals both within and outside the region contribute (through taxation) to the provision of public goods and services that benefit the region, with those not residing in the region benefiting from these policies because of presumed "altruistic" motives.[1] Treat all households within the region as a single representative individual, and let those outside the region be similarly aggregated. Suppose that each group sets its level of contribution to public goods provided within the region taking the other's contribution as given. This is most easily visualized if there is a lower-level government in the region through which the residents of the region can explicitly vary their own contributions to public goods independently of the national government, which will be assumed to ease the exposition. However, all that matters is that the political process, even if centralized, works *as if* this were the case. (In Oates' 1972 terminology, the model requires economic but not political decentralization.) Under these assumptions, we can apply the results of recent analyses of equilibrium models of public good provision by several agents (see Theodore Bergstrom et al., 1986, and Robin Boadway et al., 1989).

In conventional models of intergovernmental grants, a recipient government is treated like an individual household, and grants are viewed as a parametric change that alters the budget constraint facing the jurisdiction. Models of this type are not properly applicable, however, to situations like the present one where the recipient government and some other government are simultaneously contributing to a common public good. In the case we are considering, suppose that both the regional authority and the national government would spend resources on water projects in the absence of any ERDF grant. Then, if the grant is lump sum in form, the model predicts a reduction of national spending on water projects in the region. Because of the initial linkage between the region and the rest of the country through the national government, the grant provided to the regional authority is fully fungible with the resources of the national government, and its real impact is identical to a grant of equal size to the central government. Spending on the water project by the regional authority is predicted to remain nearly constant while national government spending on water projects in the region should fall by nearly the amount of the grant. (Since grants do not reduce spending by the regional authority on the aided function, the model implies large "flypaper effects".) Total spending from all sources on water projects in the region would thus remain roughly constant. If the grant is of a matching type, it will be offset by even greater reductions in national government spending on water projects, and it will be accompanied by a larger increase in spending by the regional authority.

---

[1] To avoid semantic confusion, note that this "altruism" need not be inconsistent with a fairly narrow conception of self-interest. Self-interested individuals may wish to commit the public sector to interventionist policy rules in order to provide social insurance. (One is reminded of H. L. Mencken's remark that "altruism, when analyzed, always turns out to be self-interest in a long-tailed coat.")

If the national government does not provide resources directly for water projects in the region, but does provide funding for, say, agricultural programs, the effects are analogous. If agricultural projects are highly substitutable for water projects, then the essential fungibility of resources is preserved, and the only difference is that the national government responds by cutting back its spending on agriculture.

Thus, the real distribution of gains from regional grants may be quite different from the apparent distribution. When a region is the beneficiary of public goods that are provided partly at the expense of other households in the country, an EEC grant may lead to reductions in support from the national government. As an example, the real beneficiaries of EEC grants to an impoverished region (or industry, etc.) in Scotland might in the end be the average U.K. taxpayer. Indeed, the painfully negotiated reductions in the U.K. contribution to the EEC budget reflect precisely this sort of logic, since it links the U.K. contribution to EEC expenditures on behalf of the U.K. This implies a willingness by the U.K. to increase its contribution to the EEC budget if it receives enough of those contributions back in the form of ERDF, ESF, or EAGGF benefits, but not otherwise, testifying to the political fungibility of resources obtained from the structural funds with those in the hands of the Chancellor of the Exchequer.

This analysis puts the structural funds in quite a different perspective than would be the case if the regions (or industries) were viewed as fiscally unconnected with other areas of the countries in which they are located. It indicates that the expenditure and revenue sides of the EEC budget are linked, and suggests the importance of thinking about the effect of the structural funds on the net fiscal position of the member states in addition to their purported functions of promoting economic development and employment in rural, declining, and poor regions and industries. Such a perspective is consistent with the approach taken by Mervyn King (undated) and Tommaso Padoa-Schioppa et al. (1987), who emphasize the need for "equitable" determination of contributions to the EEC budget, defined in terms of a country's contributions *net* of EEC expenditures in the country. As long as the programs supported by the EEC are also important functions of the member states, high substitutability between structural fund allocations and contributions must be expected. This implies that the welfare effects of these programs cannot be properly assessed in an institutional vacuum that treats other member state public sector programs as exogenously determined. Furthermore, the political pressures and constituencies that shape EEC policy are presumably highly dependent on the real distributions of gains and losses from them.

B. *Migration Spillovers*

Migration of factors of production creates another form of linkage among countries in the EEC. The nature of labor migration, and the policy issues that it poses, are likely to vary over time and space. Since migration barriers of all kinds are falling over time, these issues can be expected to grow in importance.[2] In particular, as has been emphasized in the literature, redistributional activities in one jurisdiction can produce externalities for others when migration occurs. Here I use a model that seems well suited to the analysis of redistributive activities when several jurisdictions are linked through an integrated market for some but not all types of labor. This model can be used to represent a world in which some household types find migration very easy and others find it quite difficult, which may be a particularly useful characterization of the EEC situation of partial but perhaps still limited integration of the overall labor market.

Specifically, consider a simple model where households are divided into two groups: those who are costlessly mobile and those who are completely immobile.[3] Mobile households choose locations so as to maximize income inclusive (or net) of any subsidies (or taxes). They are paid wages equal to their marginal products wherever they reside. As a condition of equilibrium, $w_i + z_i = c$ for any country i that contains mobile households, where w_i is the wage (marginal product) of mobile households in country i, z_i is a lump sum subsidy (or tax, if negative) paid to (imposed on) these households in country i, and c is the common level of net income for mobile households in other countries. All income in each country net of the wages of mobile workers accrues to the immobile households residing there. These households are assumed to control the political process in each country and to care about the welfare of the mobile households (i.e., c enters their utility functions). (As examples, think of the immobile households as rich and the mobile households as poor, or as old and young.) Because immobile households do care about the welfare of the mobile, they will either provide them with subsidies or at least will not tax away all of their income.

The important point is that the equilibrium value of c is influenced by policy choices in all countries. Subsidies to mobile workers in one country attract more of them and benefit households in other countries because of the increase in c and through a reduction in the fiscal burden of supporting mobile households. This positive externality can result in inefficiency. If each country carries out subsidies to the point that is optimal from the viewpoint of its own immobile residents, taking as given the policy of other countries, there will be underprovision of benefits to mobile households. A matching grant from the EEC would help to correct this inefficiency. In the special case where there are n countries with identical preferences, endowments, and identical technologies, the matching rate that would correctly internalize the externalities associated with support for mobile households satisfies

$$s = 1\Big/\left(1 + \frac{1}{\varepsilon}\frac{n}{n-1}\frac{w}{z}\right) \tag{1}$$

where ε is the common elasticity of demand for the labor of mobile households (defined to be positive) in each country and w and z are the common wage and subsidy rates in the symmetric equilibrium. Provided that $z > 0$, the optimal subsidy rate s lies between 0 and 1. (This formula can be used to derive illustrative calculations of the optimal subsidy.)

---

[2] It is of interest to note that residents of East Germany need only cross into West Germany to claim citizenship in the BRD, and with it the right to EEC passports. The EEC labor market has thus effectively expanded, creating a rather odd situation where East Germany is not a member state of the EEC, but its residents can claim the rights to which residents of any member state are entitled, provided only that they migrate to the West.

[3] This model is discussed in more detail in my 1990 paper. The framework of the model is closely related to that found in Boadway and myself (forthcoming). The division into mobile and immobile household types parallels my 1983 article.

The formal analysis of this model is essentially the same when the mobile group is taxed rather than subsidized by immobile households.[4] Thus, it could be applied to the analysis of taxation of wealthy mobile households by relatively immobile poorer households. In this case, there is a beneficial externality rather than a negative externality associated with migration: a taxpaying household entering a country adds a fiscal surplus rather than a fiscal burden. There is thus an incentive to tax mobile households too lightly, or to underspend on public goods, which could be offset by appropriate EEC policies.

This analysis suggests that the externalities associated with redistribution, either through income support systems or through taxation, will be exacerbated as labor markets become increasingly integrated. One of the objectives of integration is to facilitate the flow of resources, including labor, to locations where their value is high. Spatial arbitrage is inherent in this process, and in the context of labor markets this implies a tendency toward more uniform net incomes for workers. By its nature, this process implies that the effects of redistributive policies will spill across national boundaries and will be felt throughout the relevant market areas. We thus expect to find increasing pressure on EEC member states to restrict their redistributive policies, as the benefits associated with redistribution at the national level fall and the costs increase.

It is clear from the official literature of the EEC (including the Single Europe Act itself) that high priority is attached to insuring that no major groups suffer serious harm as integration proceeds. To the extent that national redistributional policies come under pressure from increasing labor market integration, one is thus likely to see expansion of redistributive policies at the EEC level—as, for example, in the doubling of the budgets for the ERDF and the ESF in the run-up to 1992.

II. Conclusion

Jacques Delors, President of the European Commission, has written that, "[a]pplying the logic of Musgrave,...a more effective allocation of resources at [the] Community level...calls for a strengthening of the redistribution function" (1987, p. *vi*). The models considered in this paper highlight different aspects of the political economy of EEC budget policy related to redistribution. The first has emphasized the potential fungibility of structural fund expenditures within member states arising from their internal political linkages, focusing attention on net fiscal transfers among member states. The second has emphasized that increasing market linkages among countries resulting from economic integration may limit the amount of redistribution undertaken at the national level in the EEC, and that this may create pressures for increased redistribution at the EEC level.

As the budgetary difficulties associated with the U.K. contributions have shown, these two considerations are not independent of one another. A significant barrier to increased redistributive expenditure at the EEC level is a concern on the part of member states that the benefits from these expenditures may be captured by other countries. The limitation on U.K. contributions is *ad hoc* in nature, and leaves the general problem unresolved. Broadly speaking, what is needed is a budgeting mechanism that provides for burden sharing that is "equitable" in that it constrains expenditures to bear a reasonably close relationship to national contributions, while providing for expenditures that promote activities generating substantial spillovers. The recent budgetary accords appear to represent progress in this direction, but budgetary issues are likely to be reopened as the EEC adapts to a rapidly changing political and economic environment.

---

[4] What matters for locational incentives is the totality of all resident-based fiscal instruments, including capital taxes assessed at the individual level as well as taxes on wage income. Paul Courant (1987) discusses the issue of tax reform in several European countries, noting that marginal tax rates for individual income taxes have recently been falling.

REFERENCES

Bergstrom, Theodore C., Blum, Lawrence and Varian, Hal R., "On the Private Provision of Public Goods," *Journal of Public Economics*, February 1986, *29*, 25–49.

Boadway, Robin W., Pestieau, Pierre and Wildasin, David, "Tax-Transfer Policies and the Voluntary Provision of Public Goods," *Journal of Public Economics*, July 1989, *39*, 157–176.

______ **and Wildasin, David E.,** "Optimal Tax-Subsidy Policies for Industrial Adjustment to Uncertain Shocks," *Oxford Economic Papers*, forthcoming.

Courant, Paul N., "Fiscal Policy and European Economic Growth," in Robert Z. Lawrence and Charles L. Schultze, eds., *Barriers to European Growth: A Transatlantic View*, Washington: Brookings, 1987.

Delors, Jacques, "Preface," in T. Padoa-Schioppa et al., eds., *Efficiency, Stability, and Equity*, New York: Oxford University Press, 1987.

King, Mervyn, "Reforming the EEC Budget," unpublished, undated.

Musgrave, Richard A., "Theories of Fiscal Federalism," *Public Finance / Finances Publiques*, No. 4, 1969, *24*, 521–32.

Oates, Wallace E., "The Theory of Public Finance in a Federal System," *Canadian Journal of Economics*, February 1968, *1*, 37–54.

______, *Fiscal Federalism*, New York: Harcourt Brace Jovanovich, 1972.

Padoa-Schioppa, Tommaso et al., *Efficiency, Stability, and Equity*, New York: Oxford University Press, 1987.

Wildasin, David E., "The Welfare Effects of Intergovernmental Grants in an Economy with Independent Jurisdictions," *Journal of Urban Economics*, March 1983, *13*, 147–64.

______, "Income Redistribution with a Common Labor Market," unpublished 1990.

U.S. Federal Budget Deficits and Gramm-Rudman-Hollings

By EDWARD M. GRAMLICH*

The U.S. federal budget deficits averaged about 1 percent of GNP in the 1960s, about 2 percent of GNP in the 1970s, and then rose to a peak of 5.5 percent of GNP in the four bad years of fiscal 1983 to fiscal 1986. By fiscal 1989, they had fallen back to 2.9 percent of GNP, and under present legislation they are now projected to drop to about 1.5 percent of GNP by 1994.

When deficits were at their peak, before the start of the budget bargain for fiscal 1987 (the first budget to show any sizeable deficit reduction), the country passed the Gramm-Rudman-Hollings (GRH) deficit control law. This law has been revised and amended over time, but it is still operating in roughly its initial form. Superficially, the numbers given above suggest that it must be working pretty well. Moreover, in many ways GRH was written to respond to the standard criticisms of deficit constraint legislation: it did not call for abrupt cuts in deficits but rather gradual declines, it was legislative and flexible rather than constitutional and rigid. Yet just as the country is on the verge of finally getting its deficits under control, the legislation is coming under vicious attack, including even that of one of its sponsors, Senator Hollings. Why should such an apparently successful piece of legislation be so widely scorned?

In this paper I take on this question. I first illustrate how GRH was intended to operate with a simple indifference curve model, and use this model to see what kinds of changes should have occurred in budget bargains and spending patterns. Then I look for these changes by comparing numbers from the pre-GRH era (1983–86) with those from the post-GRH era (1987–89). I also identify other sources of improvement in budgetary position from the earlier to the later period. This numerical comparison suggests that the direct effects of GRH played a very minor role in the improvement in the U.S. budgetary position, though it seems likely that difficult to evaluate indirect effects were much more significant.

I. Overall and Primary Deficits

Analyzing budget deficits is complicated by the fact that one component of spending is interest payments, that in turn depend on past deficits. It is easier to abstract from macroeconomic and dynamic considerations, and to identify the impact of current discretionary policy on deficits and future debt burdens, by splitting off these interest payments and focusing on what is known as the primary deficit. This is done through the following nominal identity for the change in the outstanding debt (dD):

$$dD = S - T + (r + g)D, \tag{1}$$

where S refers to program spending, T to taxes, $(S - T)$ to the primary deficit, r to the real interest rate, g to the expected rate of inflation, and $(r + g)D$ to nominal interest payments. Now letting Y refer to nominal GNP, n to the growth rate of real GNP, and differentiating

$$d(D/Y) = (S - T)/Y + (r - n)D/Y. \tag{2}$$

The dependent variable here is the change in the debt burden ratio, the change in the ratio of the stock of outstanding debt to GNP, and $(S - T)/Y$ is called the primary deficit ratio.

Table 1 gives numbers for some distinguishable epochs in the United States since

---

*University of Michigan, Ann Arbor, MI 48109. Paul Courant, Robert Inman, Rudolph Penner, and Paul Van de Water have made helpful comments.

TABLE 1

Year	$(S-T)/Y$	dD/Y	$d(D/Y)$
1962–70	−0.5	0.9	−15.8
1971–80	0.8	2.3	−2.4
1981–82	0.9	3.4	7.7
1983–86	2.4	5.5	8.6
1987	0.3	3.4	0.3
1988	0.1	3.2	−0.2
1989	−0.4	2.9	−0.5
1990–94	−1.0	2.0	−2.8

1960.[1] The first column is the primary deficit ratio averaged over the relevant fiscal years, $(S-T)/Y$; the second is the overall deficit ratio averaged over the relevant fiscal years, dD/Y; and the third is the change in the debt burden ratio from the start to the end of the relevant period, $d(D/Y)$, all in percentage points.

In the 1960s, the United States ran primary surpluses and saw its debt burden drop sharply. The primary deficit ratio rose in the 1970s, and the debt burden ratio stopped dropping sharply. In the bad years of 1983–86, the primary deficit ratio shot up and the debt burden ratio did likewise. Then there was a sharp improvement in the primary deficit ratio beginning in fiscal 1987, and the country is now following a trend where the debt burden ratio is once again declining. The key recent break in the pattern, then, was in 1987 when the primary deficit ratio not only dropped sharply, but dropped enough to change the trend of the debt burden ratio from positive to negative. This key break came just as the GRH law was instituted. Was this coincidental, or did GRH cause the turnaround?

II. A Model of GRH

The budget process for any fiscal year begins about eight months before the start of that year when the president submits his budget proposals to Congress. The Congress is then supposed to enact its own budget in timely fashion, and then pass tax and appropriation legislation consistent with this budget before the start of the fiscal year. The timing in this scenario almost always gets screwed up as the parties make their political bargains, and before GRH the numbers got screwed up too, as witnessed by the high primary deficits in the 1983–86 period.

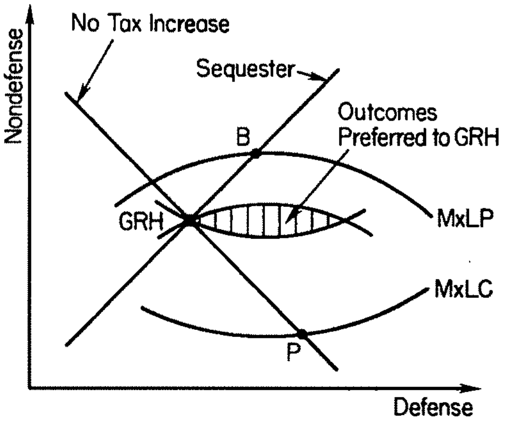


FIGURE 1

Under the present version of the GRH law, both the president's and the congressional budget must be consistent with a predetermined deficit target. If, by the start of the fiscal year, Office of Management and Budget (OMB) calculates that the deficit will be more than \$10 billion above this target, primary spending is "sequestered" to get down to the target. While some types of spending are exempt from the sequester (Social Security, low-income programs, contractual obligations), the sequester percentage is applied at one common rate for all nonexempt defense budget authority, and another common rate for all nonexempt nondefense budget authority, with these rates worked out so that the cut in defense and nondefense spending will be shared equally and will be large enough to hit the deficit target.

A simple diagrammatic model of how it was supposed to work circa fiscal 1987 is shown in Figure 1. Assume we have a president who favors high levels of defense spending, is inalterably opposed to tax increases, and is willing to cut nondefense spending whatever amount it takes to hit the deficit target. This president's preferred point is labelled P on the figure, on the No Tax Increase line. Then we have a more liberal Congress that favors slightly lower defense spending and much higher nondefense spending, at point B. Point B is labelled such because it also refers to the baseline spending level, the level that would exist under all present tax and entitlement legislation and indexed levels of all appropriated primary spending. While it is certainly not necessary to equate Congress' preferred point with the baseline spending point, most members of Congress responded to the budget crisis in roughly this way—they could not bear to make programmatic cuts in real levels of either defense or nondefense spending, their only cut proposals were elaborate forms of nominal spending freezes, and while nobody ever said so publicly, there was a strong suspicion that the Congress was not as opposed to tax increases as was the president.

---

[1] These and all other numbers throughout the paper are taken from various Congressional Budget Office (CBO) annual reports and analyses of the President's Budget. All budgetary members are divided by actual GNP, hence making the GNP shares slightly different than those published at the time. All future numbers are based on the most recent CBO projections of the impact of present legislation.

Suppose there were no agreed on budget. Before GRH, there would be a continuing resolution enacted at the eleventh hour that would place spending roughly at B. With GRH, there is a sequester that cuts defense and nondefense spending equally, moving in on the 45° Sequester line to the No Tax Increase line, at point GRH.

The diagram also assumes that both the president and the Congress have convex preferences for defense spending, nondefense spending, and private goods (avoidance of high taxes). At the president's preferred point P, Congress is on the indifference curve corresponding to its maximum loss, MxLC. At the baseline point B, the president is on the indifference curve corresponding to his maximum loss, MxLP. At the GRH point, neither party is on a preferred point, but both improve from their maximum loss level. They could do still better in the shaded area. That property was well-recognized in the drafting of the GRH law—the sequester was commonly referred to as the club in the closet that would force parties to bargain in a serious way.

How to get to this shaded area? The obvious way is to pass a tax increase. The reasons why politicians might not want to propose tax increases are well-known, but there are sneaky ways. Fiscal 1987, for example, was the first year of the new Tax Reform Act (TRA), which while revenue-neutral over a five-year period, actually gained revenue in the first year. There are also a number of less publicized pollution charges, user fees, and telephone excises that can be quietly raised or extended.

Another way is to fudge the deficit estimate, and there turn out to be many ways to do that too. The most obvious is for OMB to upgrade its economic assumptions and revenue forecasts. Another is to defer expenses or put them off-budget. Another is to move paydays forward into the present fiscal year or backward into the next fiscal year. Another is to sell loans. Every year CBO tabulates these "one-shot" and permanent deficit cuts, and the one-shot cuts have always been substantial.

Does that mean that the whole enterprise is a sham? Not necessarily. Although the matter should be studied more carefully, most one-shot cuts seem to have the property that they lower the measured deficit in one year but do not much change the future path of deficits.[2] If these fudges are simply ignored, and spending measured at its permanent level, the final outcome will be somewhere in the shaded area. This will generally, but not always, entail cuts in permanent primary spending as compared to the default level B (nondefense spending is clearly less, defense spending usually less).

There have now been three fiscal years completed under GRH. In all three years, the final outcome, as measured by the bud-

---

[2] The case is easiest to see for loan sales, the most common fudge. When the government sells a loan, it receives in cash the present value of all future loan payments. The current cash deficit drops by this amount. This cash is used to pay off outstanding debt, and reduce all future interest outflows the same amount in present value terms as the reduced loan payment inflows. Putting things off-budget is also not as magical as it seems because usually the interest carrying costs of the off-budget enterprise must remain on-budget. And for defense spending, while the common device of postponing purchases typically raises the cost of particular goods, the bargaining model used here suggests that defense postponements also lower aggregate future defense purchases.

TABLE 2—DEFENSE SPENDING

Year	Pres.	Base.	Res.
1983	6.7	6.4	6.4
1984	6.7	6.6	6.5
1985	6.9	6.6	6.7
1986	6.8	6.8	6.4
1987	6.7	6.4	6.3
1988	6.2	6.1	6.1
1989	5.7	5.7	5.7

get resolution value of permanent primary spending, was outside the No Tax Increase line but, according to my calculations, inside the shaded area. Each time spending was cut as compared to the baseline, but never was it cut as a result of an explicit GRH sequester. Each time the avoidance was accomplished with a liberal dosage of forecasting wizardry and creative fudges, and this generated much skepticism about the whole GRH process. But the real test of GRH as a deficit-cutting measure is not whether there was fudging, but whether, in the end, permanent budget resolution spending levels were lower in the GRH era than before. It is this question I try to answer in the next sections.

III. Defense Spending

The typical pattern for defense spending in the 1980s was for the president to ask for a big amount, typically more than the baseline level. Congress would usually favor a smaller amount. One might expect the overall outcome in the budget resolution to be somewhere between the wishes of the two parties, but in fact the final outcome was very close to the baseline level, both before and after GRH. The numbers, as a percent of GNP, are shown in Table 2.

It is not possible to generate more than seven annual observations with the set of political views represented in my GRH model. If any test with seven observations can be convincing, the impact of GRH can be tested formally with the time-series regression:

$$(3)\quad R-B=c_1(P-B)+c_2Z(P-B)$$

where R gives the budget resolution value for defense spending, B the baseline level of defense spending and P the president's proposal. The c_1 parameter measures the relative weight of the president's preferences in shaping budget resolution defense spending before GRH ($c_1=1$ means the president is all-powerful, $c_1=0$ means he has no impact). Z is a dummy variable that is zero in the four pre-GRH years (1983–86) and one thereafter: hence the president's weight in the post-GRH period is (c_1+c_2).

Since baseline defense spending is always below the president's proposed level, GRH cuts spending if it lowers the weight of the president in shaping spending. Unfortunately what happens when the above model is fit is that c_1 is negative and insignificant, indicating that the president's desire to raise spending above the baseline was ignored even before GRH. The parameter c_2 was also negative, indicating that the president's weight is reduced, but it is not clear how seriously to take this coefficient given that the president's weight is already outside the plausible range. To the extent that one can conduct a formal test, it indicates that GRH has not seemed to change budget bargaining on defense spending. The best predictor of budget resolution defense spending before and after GRH is simply the baseline level of spending.

But there is another story on defense spending, and here there might be a more subtle impact of GRH. As mentioned, baseline spending is determined by projecting present legislation, using real values where there is no legislation. Before GRH, present legislation was taken to mean some earlier budget agreement for real growth in defense spending—this is why baseline defense spending rose as a share of GNP in the 1983–86 period. GRH supplanted this budget agreement with a precise definition of the sequester base for defense spending, essentially the present real level—this is why baseline defense spending fell as a share of GNP in the 1987–89 period. The fall in baseline defense spending then seems to have pulled down budget resolution defense spending from a 1983–86 average of 6.5 percent of GNP to a 1989 value of 5.7 percent of GNP, a change of 0.8 percent of

TABLE 3—NONDEFENSE SPENDING

Year	Pres.	Base.	Res.
1983	13.8	15.2	14.3
1984	13.6	14.3	14.2
1985	13.8	13.9	13.7
1986	13.1	13.8	13.4
1987	12.9	13.5	13.0
1988	12.6	13.4	12.9
1989	12.8	13.0	12.7

GNP. Whether this change in the way baseline defense spending was determined was a result of GRH or simply reflected a political consensus that defense spending had become large enough is unclear. What is clear is that this change did seem to be one factor in the decline of primary budget deficits.

IV. Nondefense Spending

For nondefense spending, roles were reversed—the president wanted to cut and the Congress wanted to maintain or increase. This time the budget resolution level was nicely in the middle, with spending cut below the baseline but generally not as much as the President wanted. The numbers, again as a percent of GNP, are in Table 3.

Fitting the model described in (3) for nondefense spending leads to a statistically significant estimate of $c_1 = .55$ and a borderline insignificant estimate of $c_2 = .21$. The former means that the president and Congress have approximately equal weight in shaping nondefense spending, and the latter indicates that the president's desires to cut have received slightly more weight in the post-GRH period. Whether this change in bargaining weights can be attributed to GRH, it does not count for much—the change from the 1983–86 period to 1989 only lowers the primary deficit by 0.1 percent of GNP.

Again the big news from the numbers is not anything about the budget bargain but the sheer decline in all three sets of numbers—baseline nondefense spending, the president's proposals, and the budget resolution outcomes. This decline is again due to nondefense baseline drift, the fact that entitlement and other spending levels are set in real terms and GNP is rising in real terms. Comparing the 1983–86 average with the 1989 value for budget resolution nondefense spending, this drift has accounted for a drop of 1.2 percent of GNP in the primary budget deficit.

V. Overall Assessment

These calculations were for the defense and nondefense components of primary spending in the budget resolution. To get to the impact on primary deficits, two further steps are necessary:

There is normally a postresolution spending error that brings actual primary spending above the level in the budget resolution. For 1983–86, this postresolution error raised actual spending above budget resolution spending levels by 0.3 percent of GNP; for 1987–88 the postresolution error was slightly less, for 1989 it was 0.5 percent of GNP.

While there has been little discussion of it, 1989 actually saw a tax increase as compared to the 1983–86 average. In the pre-GRH period, total taxes averaged 18.3 percent of GNP; in 1989, they were 19.3 percent of GNP. A rise in Social Security taxes from the 1983 compromise accounted for about half of this change and the rise in corporate taxes from TRA the other half. Individual income taxes were supposed to have fallen with TRA, but they too still account for a larger percent of GNP than in the mid-1980s.

To bring all these numbers together, earlier we saw that the main factor leading to the recent decline in budget deficits was the 2.8 percentage point drop in the primary deficit in 1989, as compared to the 1983–86 average. Putting all other factors on the same basis (1989 as compared with the 1983–86 average), the drop can be decomposed as follows: Calculation of defense base 0.8; Nondefense baseline drift 1.2; Change in nondefense bargaining weights 0.1; Change in postresolution error −0.2; and Higher taxes 1.0.

The GRH law played no direct role in two factors that account for more than three-fourths of the change—the nondefense baseline drift and the higher taxes. It could have played a role in the new way to calculate the defense baseline, but it is unlikely that such an important shift would be left to GRH accountants—no doubt the redefinition of the defense baseline at least partly reflected the changing politics of defense spending. The only real change that follows the bargaining theory laid out earlier is that nondefense spending in the budget resolution is now slightly below what would have been predicted from a regression fit to earlier data. Pretty small potatoes.

From this standpoint, then, the fact that GRH was instituted just as primary deficits were dropping seems largely coincidental—defense spending had lost its political constituency anyway, nondefense spending was drifting down anyway, and taxes were going up anyway. Budget bargaining, or any changes in the process due to GRH, seemed to have little to do with the improvement.

But this story is entirely in terms of identifiable numbers. Those who are still fond of GRH point to a deeper nonquantitative argument—that GRH fundamentally changes the way Congress does business in all sorts of areas. On the spending side, this argument says either that spending would not have drifted down without GRH, or that new spending programs would have been adopted. On the tax side, the argument is that GRH has altered the way that deals were struck on TRA, as compared with earlier tax bills. Rather than permitting legislators to compete in giving away tax benefits, the GRH rules enforced in the bargaining on TRA forced all giveaway amendments to be deficit neutral, which took all the fun out of giveaways. A related experience is with catastrophic health care for the aged. This bill passed on a deficit-neutral basis too, and the unpopularity of the surtax imposed to pay for its benefits ultimately brought down a bill that probably would still exist had budget bargaining been done in the old way.

On the other side, those who are opposed to GRH point to the breakdown of discipline when numbers are fudged, the efficiency costs of continual rewriting of the tax base to hit short-term targets, the further efficiency cost of irrational cuts to meet GRH targets as opposed to sensible long term reductions, and the fetish on short-term targets in general. Nobody is against sensible long-run reductions, but where were they before GRH? Nobody is for fudging or a short-term fetish, but the model above showed that budget fudging would not necessarily worsen the permanent budget situation as compared with a no-constraint regime. And the numbers themselves show continuing downward drifts in baseline spending as GRH remains in effect. There must have been some permanent cuts.

Hence, at bottom, it is a close call for all the usual reasons. The numbers are tricky to interpret, indirect effects are impossible to evaluate and larger than direct effects, and significant changes were occurring at the same time. But to come back to the starting point of the paper, the country has finally gotten much better hold of its budget deficits. Primary deficit ratios are now dropping at a noticeable rate and the debt burden ratio is too. So even though it is difficult to pinpoint explicit gains from GRH, it is also risky to change a law that seems to be working pretty well.

Public Debts and Fiscal Politics: How to Decide?

By Robert P. Inman*

The phenomenal growth in our public deficits over the past twenty years is a matter of public and professional concern (see Symposium, 1989). While this concern may indeed be well founded (particularly given current deficit levels), public deficits are not always bad. Certainly in times of deep recessions, short-term deficit financing can stimulate aggregate demand, increase national income, and reduce unemployment in a Keynesian fashion (see Richard Startz, 1989).

Second, long-term deficit financing may be needed to sustain a long-term path of optimal consumption (see Peter Diamond, 1965). Third, public deficits (including pay-as-you-go Social Security) can offer welfare gains for significant subsets of consumers who have been liquidity constrained (R. Glenn Hubbard and Kenneth Judd, 1986), who have been unable to purchase indexed private annuities (Alan Blinder, 1988), or who have been unable to sufficiently diversify their investment portfolios away from human capital (Robert Merton, 1983).

Finally, Robert Barro (1979) has argued that public deficits may be an important policy instrument to insure intertemporal welfare maximization when public expenditures are stochastic and public taxes are economically inefficient; taxes can be "smoothed" to reduce the lifetime excess burden of public financing. The concern today is not that we have deficits, but rather that we may be overdoing a good thing. If this is the issue, then we need to ask: *Why*, and *what can be done about it*? This is my agenda here.

I. The Political Economy of Recent Deficits

Why have federal deficits increased so significantly over the past twenty years? The answer suggested here is institutional. Following the "democratization" of Congress in the late 1960s and early 1970s, the decision-making structure and the political incentives for fiscal policymaking within Congress changed to favor local constituent services first and the national interest second. In this new political environment, congressional members favor targeted spending, targeted tax breaks, and deficit financing, all enacted through omnibus spending and tax bills. Unless checked by a strong president or congressional leaders with an explicit commitment to balanced budgets, deficit financing will remain the norm.

Two sequences of events converged to produce this important shift in the structure of congressional fiscal decision making. First, American voters were becoming better educated and better informed. Old political party labels were no longer sufficient to ensure voter loyalty (see Norman Nie et al., 1979). Second, as voters became more independent, so too did their congressional representatives. With the disappearance of strong congressional leaders through retirement or political mishap followed by the democratization of the congressional committee system in the early 1970s (see Norman Ornstein, 1975), the old style of centralized, party-based policymaking gave way to a new world of individual deal making. Within this new political environment, to get anything approved often meant approval for everything. The final result is decentralized fiscal policymaking through omnibus legislation (see Barry Weingast and William Marshall, 1988).

With congressional fiscal policy now being decided in large measure by the aggregation of representatives' demands, understanding aggregate fiscal policy requires a closer look at the preferences and constraints of the members themselves. Weingast et al. (1981) have provided such an analysis; an extension of their model to include deficit financing is revealing. With a focus on local constituents, each member is interested in maximizing service benefits and the after-federal-tax income received by the voters of the state or district. Voters' after-tax income (Y_i) in a typical district i is defined as their pre-tax incomes (I_i) minus their tax share ($\phi_i \simeq$ the district's share of national income $= I_i/\Sigma I_i$) of all federal government *project* spending plus interest on prior debt ($X_i + \Sigma_{j\neq i} X_j + rD$) minus their tax share (ϕ_i) of all federal government *tax* expenditures or "loopholes" ($T_i + \Sigma_{j\neq i} T_j$) plus the direct dollar benefits of the tax expenditures received by the district's voters (T_i) plus their tax share (ϕ_i) of any (present value) tax savings made possible through federal deficit financing (d) in the current budget period ($[1-\rho_i]d$), where ρ_i is the fraction of the current deficit's repayment which cannot be avoided by the residents of district i:

---

*University of Pennsylvania, Philadelphia, PA 19104 and National Bureau of Economic Research, Cambridge, MA 02138. The comments of Paul Courant, Judith Farnbach, and John Yinger are very much appreciated as is the financial support of the NSF (SES87-18768). This paper is part of a larger project with Daniel Rubinfeld on the design of federalist fiscal institutions.

$$(1)\quad Y_i = I_i + T_i - \phi_i T_i - \phi_i X_i - \phi_i\left\{\sum_{j\neq i} X_j + \sum_{j\neq i} T_j + rD\right\} + \phi_i\{[1-\rho_i]\,d\}.$$

When congressional budgeting is decentralized, the budget costs imposed by districts other than one's own ($\phi_i\{\Sigma_{j\neq i} X_j + \Sigma_{j\neq i} T_j + rD\}$) are viewed as given by the representative of each individual district i. In effect, the individual members are playing Nash strategies against each other within the congressional budget game; the equilibrium aggregate budget will be the summation of the equilibrium requests of the individual members. The last term in the constraint is the share of current deficit financing (d) avoided by the residents of district i. If $\rho_i = 1$, current taxpayers in district i will bear their full share ($=\phi_i$) of all future tax costs associated with d. Here deficit financing offers no net benefit (but imposes no burden either) for the residents of district i. When ρ_i is less than 1, however (for example, if current voters in district i expect to die, leaving no intended bequests, before all the principal and interest of d falls due), then current borrowing will be a transfer to the current taxpayers in district i. "Rolled-over" borrowing and a growing stock of federal debt offers this possibility.

This after-federal-tax income constraint reveals two important facts about decentralized congressional budgeting. First, each representative considers only the direct costs of his fiscal choices on his constituents and is led to demand inefficiently large project spending (X_i) and district-specific tax benefits (T_i). This occurs because the marginal costs to district residents are below the social costs of the project or transfer. Second, there are no positive incentives for any individual representative to control new government borrowing (d). The incentive (if any) is to expand deficit financing if $\rho_i < 1$.

When seen in this light, the congressional budget process looks much like the individual exploitation of a common pool resource, where that resource is the current and future national tax base. As with all common pool resource problems, the individual incentives for fiscal inefficiency and deficit financing can be controlled, but only if there is a centralizing decision-making structure that can coordinate and enforce the individual representative budget requests into a more rational budget for the whole. Unfortunately, the democratization of the congressional committee structure seriously undermined this possibility. In contrast, the presidency through its power to formulate the initial budget proposal (in effect, to set the agenda for congressional decisions) may be able to provide the needed centralizing budget structure. Yet to be successful, the president must act like, and have the power of, a true agenda setter. He must present the Congress with a "take-it-or-leave-it" budget, and then have the ability to punish those members who do not conform.

Ronald Reagan's early budgets, fashioned by OMB Director David Stockman, were presented in just this manner with the intended effects. Real defense spending was significantly increased, while "controllable" domestic spending was significantly reduced (see my 1988 paper). On the tax side, the President's Economic Recovery Tax Act of 1981 (ERTA) delivered an across-the-board reduction in all rates of 25 percent over the subsequent three years (see Charles Jacob, 1985). All of this was possible because of the president's control over a Republican Senate and his ability to mobilize citizen support to keep enough conservative House Democrats in place.

What the Reagan budgets did not control were deficits, however. Presidential preferences for defense spending and reductions in the average levels of domestic spending and taxation took precedence. These presidential preferences coupled with a congressional need for distributive budgets made the deficit explosion of the mid-1980's virtually inevitable. Since the deleted spending programs gave benefits largely on a per capita basis (see my earlier paper), ERTA's uniform reduction in a progressive rate structure required a deficit-producing tax rate reduction to insure sufficiently wide House support. Representatives from the low- to moderate-income districts had to have a tax cut sufficient to compensate their constituents for losses on the spending side. The 25 percent across-the-board cut did the job, but it also overcompensated the median and upper income districts. Thus domestic spending cuts fell short of domestic tax relief.[1] When joined with the large defense buildup, large federal deficits were the result.

This recent fiscal history suggests that perhaps deficit financing is endemic to the new, decentralized politics of Washington.[2] The solution might well be to constitutionally constrain public borrowing at the national level through an enforceable (i.e., simple) balanced budget amendment, and to then turn to the state and local fiscal sector for the setting of discretionary public deficits. There is an argument for this position, and it needs to be evaluated.

II. Should We Decentralize Deficit Policymaking?

A quick examination of the facts does suggest that state and local governments are doing a better job than Washington in managing their current accounts budgetary deficits. Deficits for the state and local sector, also measured for a "unified" public budget, show state and city governments in the aggregate to have fiscal surpluses, except for small deficits (soon paid back) in recession years.[3] On the surface, at least, state and local governments seem to be avoiding the Washington propensity to deficit finance.

There is an argument why this might be so, building upon the logic of the full information, competitive local public goods economy (see Daniel Rubinfeld, 1987, for an overview). In this economy, deficit financing can be used, but only to the extent that deficits provide real long-term benefits to resident taxpayers—for example, to "tax smooth" the long-term payments for capital investments. Deficit financing that seeks to shift the burden of public goods financing from current to future, nonresident taxpayers cannot occur. Future taxpayers can obtain relief from local debt unmatched by valued assets by demanding a fully compensating reduction in house (or business) prices from current residents when they enter the community. Importantly, this reduction in house prices will reflect *all* future burdens of the debt, *even* (in contrast to national deficits) when that debt is "rolled-over" into perpetuity. The capitalization of the burden of future deficits into land prices insures that current residents who move still bear the full burden of their local deficits. If local taxpayers control local fiscal policy, then all fiscal benefits and costs will be weighed appropriately, and optimal short- and long-term fiscal policies will result.[4]

---

[1] ERTA could have chosen smaller percentage tax cuts for the middle- and upper-income tax brackets, but that might have opened the door (even wider) to a tax loophole bill. ERTA's uniform percentage cut in rates is consistent with the uniform percentage cut in spending, Stockman's agenda-setting strategy in those early budgets.

[2] Recent work by Nouriel Roubini and Jeffrey Sachs (1989) suggests that the recent deficit experience in the United States may generalize to other countries with decentralized political structures.

[3] See Table 1 to a longer version of this paper available from the author upon request.

This extension of the usual local public goods economy to include efficient intertemporal fiscal policy rests (like its atemporal counterpart) upon the validity of the underlying assumptions, two of which seem particularly suspect here. First, the analysis assumes that residents have full information about each community's borrowings and resulting net public wealth. Unfortunately, such information is not only hard to collect, but also hard to interpret. To do justice to the task, each resident in the local economy must be an accountant, an actuary, a civil engineer, and an economist. Of course, citizens might hire experts, but when market prices reveal private information (as will occur here through the housing market), each market participant has an incentive to free ride, hoping others will provide the "facts." The likely outcome is an underprovision of information about local deficits. If so, then borrowing can benefit current voting taxpayers at the expense of nonvoting future residents. In such an economy, current residents' dominant strategy will be to borrow to protect themselves against the risk, when moving, of assuming someone else's undetected local deficits. The result will be too much local borrowing. There is good evidence that such incentives have driven the recent buildup of state pension deficits, one off-budget local deficit which is hard to measure (see my paper with David Albright, 1990).

Second, even if full information about local deficits does obtain, the intertemporal efficiency result requires that local deficit policy internalize all appropriate external benefits and costs. This is unlikely in our economy dominated by national markets and national market failures. Two of the several market failures for which public deficit policy is an appropriate policy response (Keynesian recessions and aggregate capital accumulation) involve national externalities for which local deficit policy will prove inadequate.[5] It is unlikely any one city or state has sufficient collateral to borrow what might be needed, or that its private export economy is so fully integrated that local deficits can insure national (full employment) efficiency.[6] Even so, there remains a possible incentive problem. If local deficits are known, then deficits that do not create a direct local benefit will be capitalized into lower local house prices. Unfortunately, it is just those deficits that potentially create the national employment or capital accumulation benefit. With local costs but national benefits, the incentive will be to underprovide the required deficits.

While there is a theoretical case for decentralizing our national deficit policy, it rests upon doubtful assumptions. Informational imperfections are likely to lead to excessive local deficits, while the external interdependencies of local private economies may lead to deficits which are too small. It is unlikely that these biases just cancel.

III. How Then to Decide?

We face a choice between imperfect fiscal institutions. In theory, Washington can internalize all appropriate current (Keynesian) and future (capital accumulation) period externalities associated with deficit financing, but the present decentralized decision-making structure is ill-suited to the task. Deficits are overused within Washington's redistributive fiscal politics. In theory, a competitive local public goods economy can control the redistributive misuse of deficits through the capitalization process, but such an economy may well fail to internalize all relevant externalities, and may itself be open to mismanagement when information about deficits is not public. Fortunately reforms are possible. At the central government level, institutional reforms that strengthen the hand of the president or of congressional leaders appear to offer some hope for a more coordinated, less distributive deficit policy (see Gramlich, 1990, and Rudolph Penner and Alan Abramson, 1988). At the local level, common auditing standards and public exposure of all local assets and liabilities (not just the current accounts budget) would do much to encourage improved fiscal discipline.

---

[4] There are two exceptions to this argument. First, when current residents are renters and therefore own no private assets within the community, capitalization can have no effect upon their fiscal choices. Second, capitalization is also ineffective when current residents do own property within the community but remain until death and leave no intended bequests. Their heirs will be affected by capitalization, but without intended bequests they have no say over current fiscal choices.

[5] In contrast, relaxing household liquidity constraints or providing indexed annuities or more diversified portfolios might all be done by local government borrowing. Minimizing the intertemporal excess burden of inefficient local taxes can also be achieved through local borrowing.

[6] Though Edward Gramlich (1987) has argued to the contrary.

Just as we seek to design market economies to appropriately harness economic incentives, so too must we seek to design governmental institutions to appropriately channel political incentives. Improved deficit policies will hopefully be one consequence of the effort.

REFERENCES

Barro, Robert, "On the Determination of the Public Debt," *Journal of Political Economy*, October 1979, *87*, 940–71.

Blinder, Alan, "Why is the Government in the Pension Business?," in Susan Wachter, ed., *Social Security and Private Pensions*, Lexington: D. C. Heath, 1988.

Diamond, Peter, "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, *55*, 1126–50.

Gramlich, Edward, "Subnational Fiscal Policy," *Perspectives on Local Public Finance and Public Policy*, 1987, *3*, 3–27.

______, "U.S. Federal Budget Deficits and Gramm-Rudman-Hollings," *American Economic Review Proceedings*, May 1990, *80*, 75–80.

Hubbard, R. Glenn and Judd, Kenneth, "Liquidity Constraints, Fiscal Policy, and Consumption," *Brookings Papers on Economic Activity*, 1:1986, 1–51.

Inman, Robert, P., "Federal Assistance and Local Services in the United States: The Evolution of a New Federalist Fiscal Order," in Harvey Rosen, ed., *Fiscal Federalism: Quantitative Studies*, Chicago: University of Chicago Press, 1988.

______ **and Albright, David,** "Local Debt in a Federalist Public Economy: Teacher Pensions as a Case Study," mimeo., University of Pennsylvania, 1990.

Jacob, Charles, "Reaganomics: The Revolution in American Political Economy," *Law and Contemporary Problems*, Autumn 1985, *48*, 7–30.

Merton, Robert, "On the Role of Social Security as a Means for Efficient Risk-Sharing in an Economy Where Human Capital is Not Tradeable," in Zvi Brodie and John B. Shoven, eds., *Financial Aspects of the United States Pension System*, Chicago: University of Chicago Press, 1983.

Nie, Norman, Verba, Sidney, and Petrocik, John, *The Changing American Voter*, Cambridge: Harvard University Press, 1979.

Ornstein, Norman, *Congress in Change*, New York: Praeger, 1975.

Penner, Rudolph and Abramson, Alan J., *Broken Purse Strings: Congressional Budgeting, 1974–88*, Washington: Urban Institute, 1988.

Roubini, Nouriel and Sachs, Jeffrey D., "Political and Economic Determinants of Budget Deficits in the Industrial Democracies," *European Economic Review*, September 1989, *33*, 903–38.

Rubinfeld, Daniel, "The Economics of the Local Public Sector," in Alan Auerbach and Martin Feldstein, eds., *Handbook of Public Economics*, New York: North-Holland, 1987.

Startz, Richard, "Monopolistic Competition as a Foundation for Keynesian Macroeconomic Models," *Quarterly Journal of Economics*, November 1989, *104*, 737–52.

Weingast, Barry and Marshall, William, "The Industrial Organization of Congress; or Why Legislatures, Like Firms, Are Not Organized as Markets," *Journal of Political Economy*, February 1988, *96*, 132–63.

______, **Kenneth Shepsle, Christopher Johnsen,** "The Political Economy of Benefits and Costs: A Neoclassical Approach to Distributive Politics," *Journal of Political Economy*, April 1981, *89*, 642–64.

Symposium, "Budget Deficit," *Journal of Economic Perspectives*, Spring 1989, *3*, 17–93.

THE "NEW" GROWTH THEORY[†]

Trade, Innovation, and Growth

By GENE M. GROSSMAN AND ELHANAN HELPMAN*

Recent papers by Paul Romer (1986, 1990) and Robert Lucas (1988) have reminded us that when investment takes place in an economic environment with increasing returns to scale, the marginal product of capital need not decline over time to the level of the discount rate. Then the incentive to accumulate capital may persist indefinitely, and long-run growth in per capita income can be sustained. These astute observations have revitalized the theory of economic growth. Research attention has focused primarily on the processes of accumulation of *knowledge capital*, in part because the public good aspects of knowledge as information naturally create increasing returns to scale in many contexts. There has been an attempt to understand the determinants of long-run growth based on investment in human capital and in new technologies.

The advances in growth theory enable us to address rigorously many issues that have long been central to international economics. For example, to what extent and in what ways might international trade serve as an "engine of growth"? Do international exchanges naturally enhance the growth performance of individual trading countries? And what economic policies are especially conducive to high levels of welfare in a growing, open economy?

Growth theorists also stand to gain from recognizing how the international economic environment impinges upon the incentives that firms in specific countries have to invest in the creation of knowledge. Several features of the global economy seem especially important for understanding growth performance. First, familiar notations of comparative advantage may determine to what extent particular countries are led to specialize in the creation of knowledge and in the production of goods that make intensive use of human capital and new technologies. Second, the large scale of the world economy provides great opportunities for the exploitation of research successes and so may enhance the incentives that firms have to invest in the generation of new technologies. Third, in a world of rapid and cheap communication, ideas and information spread quickly across international borders. Countries stand to benefit from the spillovers generated by investments in knowledge in trade partner countries, but also may lose from the lack of ability to appropriate all of the benefits from their own investments. Finally, participation in international capital markets provides an expanded set of opportunities for financing investments in all forms of capital, including knowledge capital. These various aspects of the international trade environment have featured prominently in our own work on innovation and growth in the open economy. In the next section we describe common elements of our research approach. Then, in Section II, we introduce a highly simplified model of trade, knowledge accumulation, and growth, and use it to expound some of our recent findings.

I. Modeling Endogenous Innovation

Many growth theorists raised in the neoclassical, Solovian tradition took technological progress to be an exogenous and fortuitous process. Several common features distinguish recent efforts to endogenize innovation within general equilibrium models of long-run growth. Foremost among these are, first, a rigorous accounting of the resources used up in creating new knowledge, and second, explicit consideration of the profit motive that drives private investment in *R&D*. In these matters, the new theory draws on modeling approaches developed by industrial organization economists.

[†] *Discussants*: Irma Adelman, University of California-Berkeley; Richard Day, University of Southern California; Meir Kohn, Dartmouth College.

*Woodrow Wilson School, Princeton University, Princeton NJ 08544, and Tel Aviv University, Tel Aviv 69978, Israel, respectively. We are grateful to Jean Baldwin Grossman, Torsten Persson, and Ken Rogoff for suggestions and to the National Science Foundation for research support.

Industrial *R&D* may be aimed at cost reduction, product innovation, or quality improvements. Recent research has incorporated all of these forms of technological progress into analyses of long-run innovation and growth. In our own work on product innovation (1989a, b, 1990; see also Romer, 1990) we assume that an entrepreneur must develop the design for a new, differentiated product before it can be produced. This requires that resources be devoted to *R&D*. We treat *R&D* as an ordinary economic activity, specifying a technology that relates inputs (primary factors of production) to outputs (blueprints for new products). In our recent papers (1989c, d) and Grossman (1989), we model the process of quality upgrading as a set of concurrent, industry-specific patent races, each aimed at developing the next generation of product.[1] In this case, a standard production function links the input of resources to the entrepreneur's instantaneous probability of achieving a research breakthrough. In either case, the cost of *R&D* depends both on technological considerations and on market conditions, the latter because factor prices are determined as usual in the general equilibrium.

Throughout our work, we have adopted a Schumpeterian perspective. Research successes generate some limited degree of market power, and so create profit opportunities. These potential profits justify the expenditures on *R&D*. When a new product is developed that substitutes imperfectly for existing brands, the innovator can establish a market niche and charge a price above marginal cost in the ensuing oligopolistic competition. Similarly, when an existing product is improved, the new industry leader can price above the cost of production and still find consumers willing to buy his superior, state-of-the-art product. In either case, the innovator earns a stream of profits that may last indefinitely, or for a limited period of time.

We model the *R&D* process as being one with free entry. Entrepreneurs may establish research labs whenever the incentive to do so is present. Then, in an equilibrium with an active *R&D* sector, the expected returns to this activity must just be "normal"; that is, they must reflect the opportunity cost of capital and compensate for any undiversifiable risks. We have used this "no-arbitrage" condition to link the interest rate determined on capital markets to the rate of profit enjoyed by successful innovators, and to the probability that an existing profit stream will be eroded or eliminated due to imitation or further innovation by rivals. By adopting this perspective, we are able to investigate several channels through which international trade and trade policies may affect the incentives that a firm in some country has to engage in *R&D*. For example, the existence of foreign competitors may shorten the likely duration of any profit opportunity if foreign firms can engage in reverse engineering or target the product in question for further improvement. And a trade barrier influences rates of profit for home and foreign firms, and also alters factor prices and so the cost of *R&D* in the general equilibrium.

Our modeling of *R&D* incorporates in all cases some of the spillovers that we believe to be inherent to the process of knowledge generation. As is well known, technology bears many of the characteristics of a public good. Knowledge as a commodity is nonrivalrous; that is, several parties may put the same information to use simultaneously at no extra cost. Also, it may be difficult in many instances to define and enforce property rights perfectly, so that parties cannot exclude the use by others of the knowledge that they may have created. In particular, an industrial innovator will have difficulty preventing others from taking advantage of the more general forms of scientific and engineering knowledge that are generated in the course of developing some specific product or process. As Romer (1990) has emphasized, these spillovers may cause aggregate investment in knowledge to exhibit nondecreasing returns to scale and so allow innovation to be a sustainable process in the long run.

---

[1] This approach draws on Phillipe Aghion and Peter Howitt (1989).

The exact form that the spillovers take may vary in different applications and for different types of industrial research. We have so far adopted two different modeling approaches. In our papers on product innovation, we follow Romer (1990) in assuming that each research project generates not only a patentable blueprint for its perpetrator, but also a nonappropriable contribution to the stock of general knowledge capital. We treat knowledge capital as a (public) input into *R&D*, so that at any point in time fewer resources are needed to invent a new variety of product the greater is the state of scientific understanding. In our papers on quality upgrading, we assume that research labs can enter the race for the next generation technology even if they have not succeeded in bringing out the current generation product. Implicitly, we distinguish the knowledge needed to manufacture a good (or the legality of doing so under patent rights protection) from the knowledge needed to try to invent a better product. We assume that production know-how is private and appropriable, while improvement know-how lies in the public domain.

II. Determinants of Long-Run Growth

In order to exposit some of our findings, it proves convenient to develop a very simple, indeed almost trivial, model of trade, knowledge accumulation, and endogenous growth. The model bases growth on learning-by-doing in a single, knowledge-generating sector. It therefore neglects several important elements that are central to the approach described in Section I. In particular, there is no separate *R&D* activity, and no resources are devoted exclusively to the task of generating new knowledge. Moreover, knowledge accumulation is not guided by the profits that accrue to the owners of new or superior technologies. Still, the simple economy here shares two features with the more complex economies described in Section I. First, the growth rate is proximately determined by the equilibrium allocation of resources to a knowledge-creating activity; here the sector that generates learning-by-doing benefits plays the role of the *R&D* sector in our other papers. Second, spillovers play a critical role in making long-run growth sustainable. Due to these similarities, the simple model can serve a useful pedagogic role.

Consider then a two-sector, two-factor economy. Let the factors, land and labor, be available in fixed supplies, T and L. Output in sector i is given by a constant-returns-to-scale, neoclassical production function

$$(1) \qquad X^i = KF^i(T^i, L^i), \quad i=1,2,$$

where T^i and L^i are employments of land and labor, respectively, in sector i, and K (for "knowledge") represents the instantaneous stock of knowledge capital, a public input. Knowledge accumulates as a by-product of manufacturing experience in one of the sectors, say sector 1. These learning-by-doing benefits, which augment productivity in both sectors, are wholly external to the individual firms that generate them. We suppose that

$$(2) \qquad \dot{K} = bX^1.$$

Finally, consumers maximize any homothetic, intertemporal utility function.

Let us begin with a small economy that trades the two goods at exogenous relative prices, $p \equiv p_1/p_2$. Suppose for the moment that it enjoys no knowledge spillovers from abroad. Each small firm in sector 1 ignores its (nonappropriable) contribution to future knowledge, and so maximizes instantaneous profit. Clearly the equilibrium allocation of resources is the same as for a static, competitive economy with production functions F^i and total factor supplied T and L; that is, the marginal rate of transformation between goods 1 and 2 is set equal to p. Then $\dot{X}^i = F^i(T^i, L^i)bX^1$, and output in each sector grows at the constant rate $g = bF^1(T^1, L^1)$.

Suppose now that the supply of one of the factors were to increase. Several authors have found a positive relationship between the size of the resource base and the rate of growth. Here, the Rybczynski Theorem implies that an increase in the supply of the resource used intensively in the knowledge-generating sector speeds growth, but an increase in the supply of the resource used intensively in the production of good 2 slows growth. We found similar results in our 1989d paper and in Grossman (1989). There, *R&D* and the production of high-technology goods act like a joint activity in the free-trade equilibrium, and it is the factor intensity of the composite activity that matters for predicting the effect of factor accumulation on the rate of growth.

Next consider the effects of trade policy. Here, protection of sector 1 shifts resources into the knowledge-creating activity and so speeds the rate of growth. Protection of sector 2 has the opposite effects on resource allocation and growth. More generally, we found two distinct influences that trade policy has on the growth rate. First, protecting some sector augments the derived demand of that sector for the output of the *R&D* activity. Put differently, the return to a research success generally will rise when the sector in which the technology is applied is promoted via trade policy. But second, there will be an influence that operates through factor markets, similar to the one that applies in the simple model here. If the government protects a production sector that competes with *R&D* for resources, then the cost of *R&D* will rise and the allocation of resources to this activity may shrink. In our 1989b paper, where human capital is devoted to inventing new, nontraded, intermediate products, we found that promotion of the human capital-intensive final good via trade policy is detrimental to growth, while promotion of the labor-intensive good is conducive to growth. The former sector is a general equilibrium substitute for *R&D*, while the latter is a general equilibrium complement. In Grossman (1989), protection of the high-technology sector via trade policy causes skilled labor to shift from research to manufacturing and so retards innovation in the policy active country.

The welfare economics of our simple economy are equally transparent. The allocation of resources to sector 1 is suboptimal, because this activity generates a spillover benefit that private agents do not take into account. A first-best policy subsidizes output in this sector, while commercial policy represents a second-best government intervention. These welfare-improving policies raise the growth rate. But the optimal growth rate may fall well short of the maximal rate, and policies that cause an excessive allocation of resources to knowledge creation can reduce welfare.

The welfare analysis of a Schumpeterian economy with an explicit *R&D* activity shares some of these features, but involves further complexities. There exists in such economies a second distortion besides that due to the spillovers generated in the creation of knowledge. This distortion arises from the noncompetitive pricing of innovative products by successful entrepreneurs. It leads to an undersupply of the volume (as opposed to the number) of innovative products. We show in our 1989b paper that equilibrium growth is nonetheless too slow in our particular specification of product innovation (based on the familiar Dixit-Stiglitz formulation), but our 1989c paper demonstrates that the equilibrium allocation of resources to *R&D* may be excessive when research generates quality improvements (see also Aghion and Howitt). In the former case, an appropriate subsidy to *R&D* always raises welfare, while in the latter case a tax on *R&D* may be desirable. Moreover, in these richer economic environments, even if the growth rate is too slow, a trade policy that encourages growth may fail to provide second-best welfare benefits. A trade policy that increases the rate of growth toward the first-best level may lower welfare if it at the same time reduces the level of output of the non-competitively priced commodities (see our 1989b paper).

Equation (2) points to the potential role of international spillovers in the growth process, and suggests another mechanism by which commodity trade might influence growth. The simplest specification would make technological progress ($\dot{K}$) a function of *world* output of good 1. Then the home country would automatically enjoy the benefits of knowledge created abroad. But this specification probably is not descriptive of a large portion of the international diffusion of technology that actually takes place. Often it is necessary for local firms to invest resources in order to capture spillover benefits from abroad. In our models of the product life cycle, we suppose that products initially developed in the industrialized "North" later become candidates for (costly) imitation by the semi-industrialized "South." We use this formulation to study the growth effects of North-South trade.

Our 1989a paper provides an example where such trade must speed growth in both regions; product imitation by the South raises the North's incentive to innovate, because firms in the North earn greater profits during their period of monopoly production. Even though imitation spells the eventual end to their profit stream, the expected present value of the returns to any innovation are increased by trade. While this result does not generalize to all specifications (see our 1989c paper), it does suggest that the link between the rate of technological diffusion (as influenced, for example, by policies regarding the protection of foreign intellectual property) and the equilibrium rate of innovation and growth is more subtle than might appear at first glance.

Let us consider now a two-country world economy, each with production functions as in (1) and with technological progress in (2) depending upon world output of good 1. Suppose that the home country were to subsidize output of this good or promote the industry by means of a trade policy. Resources in this country would shift into sector 1, which *ceterus paribus* would accelerate growth in the world economy. But, in the foreign county, the opposite resource movements would take place. The contribution of the foreign country to knowledge accumulation would decline. The net effect on technological progress and hence world growth would depend upon the balance of these offsetting influences. This balance, in turn, would depend in part upon which of the two countries enjoyed comparative advantage in the production of the X^1.

In our 1990 paper and in Grossman (1989), we studied the effects of trade and industrial policies on the long-run rate of growth in the world economy. We considered policies introduced by a single country in situations where learning externalities from *R&D* are international in scope. Our findings were quite similar to those for the simple economy here. If a country with comparative advantage in undertaking *R&D* were to subsidize research, the world growth rate would accelerate. But if a similar subsidy were to be introduced by the country that is relatively better at manufacturing rather than innovating, the world growth rate *may* decline. Similarly, protective trade policy will raise long-run rates of growth when undertaken in a country with comparative disadvantage in *R&D*, but will have the opposite effect on growth when implemented by a country with comparative advantage in *R&D*. Since we find that trade protection shifts resources from research into manufacturing in the policy active country, and in the opposite direction in the policy inactive country, our findings are understandable in the light of the results for the simple model of this paper.

III. Concluding Remarks

Casual observation and more systematic empirical research suggest that countries that have adopted an outward-oriented development strategy have grown faster and achieved a higher level of economic well-being than those that have chosen a more protectionist trade stance. The evidence on the efficacy of explicit policies to promote exports, including various forms of industrial targeting, is not yet conclusive. The approach to modeling endogenous innovation and endogenous human capital formation that has been proposed here may provide a means for improving our understanding of the connection between the international trade environment including the trade policy regime and long-run growth performance. It seems clear that the less developed countries *potentially* stand the most to gain from their international relationships, since in principle these countries can draw upon the large stock of knowledge capital already accumulated in the industrialized world. But, it is equally clear from the poor growth experience of some of these countries, that the technology flows are anything but automatic. We need to learn much more about the mechanisms by which knowledge and technology diffuse across international borders (including, for example, the role of multinational corporations in this regard), and the incentives that impinge upon the equilibrium rate of technology transfer.

REFERENCES

Aghion, Phillipe and Howitt, Peter, "A Model of Growth Through Creative Destruction," MIT, mimeo, 1989.

Grossman, Gene M., "Explaining Japan's Innovation and Trade," Woodrow Wilson School Discussion Paper in Economics No. 151, Princeton University, 1989.

______ **and Helpman, Elhanan,** (1989a) "Endogenous Product Cycles," NBER Working Paper No. 2913, 1989.

______ **and** ______, (1989b) "Growth and Welfare in the Small Open Economy," NBER Working Paper No. 2970, 1989.

______ **and** ______, (1989c) "Quality Ladders and Product Cycles," Woodrow Wilson School Discussion Paper in Economics No. 152, Princeton University, 1989.

______ **and** ______, (1989d) "Quality Ladders in the Theory of Growth," NBER Working Paper No. 3099, 1989.

______ **and** ______, "Comparative Advantage and Long-Run Growth," *American Economic Review*, *80*, forthcoming 1990.

Lucas, Robert E., Jr., "On the Mechanics of Economic Development," *Journal of Monetary Economics*, January 1988, *22*, 3–22.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–37.

______, "Endogenous Technological Change," *Journal of Political Economy*, *98*, forthcoming 1990.

Why Doesn't Capital Flow from Rich to Poor Countries?

By ROBERT E. LUCAS, JR.*

The egalitarian predictions of the simplest neoclassical models of trade and growth are well known and easy to explain, as they follow from entirely standard assumptions on technology alone. Consider two countries producing the same good with the same constant returns to scale production function, relating output to homogeneous capital and labor inputs. If production per worker differs between these two countries, it must be because they have different levels of capital per worker: I have just ruled everything else out! Then the Law of Diminishing Returns implies that the marginal product of capital is higher in the less productive (i.e., in the poorer) economy. If so, then if trade in capital good is free and competitive, new investment will occur only in the poorer economy, and this will continue to be true until capital-labor ratios, and hence wages and capital returns, are equalized.

We do, of course, see some investment by wealthy countries in poorer ones, but an example with some rough numbers will help to make clear just how far the capital flows we observe fall short of the flows predicted by the theory I have just sketched. According to Robert Summers and Alan Heston (1988, Table 3, pp. 18–21), production per person in the United States is about fifteen times what it is in India. Suppose production in both these countries obeys a Cobb-Douglas-type constant returns technology with a common intercept:

$$y = Ax^{\beta}, \tag{1}$$

where y is income per worker and x is capital per worker. Then the marginal product of capital is $r = A\beta x^{\beta-1}$, in terms of capital per worker, and thus:

$$r = \beta A^{1/\beta} y^{(\beta-1)/\beta} \tag{2}$$

in terms of production per worker. Let $\beta = 0.4$ (an average of U.S. and Indian capital shares), again for both countries. Then the formula (2) implies that the marginal product of capital in India must be about $(15)^{1.5} = 58$ times the marginal product of capital in the United States.

If this model were anywhere close to being accurate, and if world capital markets were anywhere close to being free and complete, it is clear that, in the face of return differentials of this magnitude, investment goods would flow rapidly from the United States and other wealthy countries to India and other poor countries. Indeed, one would expect *no* investment to occur in the wealthy countries in the face of return differentials of this magnitude. I worked out the arithmetic for this example to make it clear that there is nothing at all delicate about this standard neoclassical prediction on capital flows. The assumptions on technology and trade conditions that give rise to this example must be drastically wrong, but exactly what is wrong with them, and what assumptions should replace them? This is a central question for economic development. I consider four candidate answers to this question.

I. Differences in Human Capital

The sample calculation in my introduction treats effective labor input per person as equal in the countries being compared, ignoring differences in labor quality or human capital per worker. The best attempt to correct measured labor inputs for differences in human capital is Anne Krueger's study (1968). Her estimates are based on data from the 1950s, but the percentage income differentials between very rich and very poor countries have not changed all that much in the last 25 years and, in any case, a rough estimate is better than none at all. Her method is to combine information on each country's mix of workers by level of education, age and sector with U.S. estimates of the way these factors affect worker productivity, as measured by relative earnings.

---

*University of Chicago, Chicago, IL 60637. I am grateful to Nancy Stokey for helpful criticism, and for research support under NSF grant no. SES-8808835.

Krueger's main results are given in her Table III (p. 653), that gives estimates of the per capita income that each of the 28 countries examined could attain, expressed as a fraction of U.S. income, if each country had the same physical capital per worker endowment as did the United States. The estimates range from around .38 (India, Indonesia, Ghana) to unity (Canada) and .84 (Israel). These numbers have the dimension of the relative human capital stocks raised to the power of labor's share, so taking the latter at .6 (as I did in my introductory example), the estimated relative human capital endowments ranged from about .2 to unity. That is, each American or Canadian worker was estimated to be the productive equivalent of about five Indians or Ghanians. (Compensation per employed civilian in the United States in 1987 was about $24,000, so this estimate implies that a typical worker from India or Ghana could earn about $4800 in the United States.)

To redo my introductory example with Krueger's estimated human capital differentials, reinterpret y in equations (1) and (2) as income per *effective* worker. Then the ratio of y in the United States to y in India becomes 3 rather than 15, and the predicted rate of return ratio becomes $(3)^{1.5} = 5$ rather than 58. This is a substantial revision, but even so, it leaves the original paradox very much alive: a factor of 5 difference in rates of return is still large enough to lead one to expect capital flows much larger than anything we observe.

If it had turned out that replacing labor with effective labor had *entirely* eliminated estimated differences in the marginal product of capital, this would have answered the question with which I began this paper, but only by replacing it with an even harder question. Under constant returns, equal capital returns implies equal wage rates for equally skilled labor, so that if there were no economic motive for capital to flow, there would be no motive for labor flows either. Yet we see immigration at maximal allowable rates and beyond from poor countries to wealthy ones. We do not want to resolve the puzzle of capital flows with a theory that predicts, contrary to the evidence provided by millions of Mexicans, that Mexican workers can earn equal wages in the United States and in Mexico.

II. External Benefits of Human Capital

Obviously, we could resolve the puzzle of the inadequacy of capital flows at any time by *assuming* that marginal products of capital are equalized, and using equation (2) and the estimated income differential to estimate the relative levels of the intercept parameter A (often called the level of technology) in the two countries being compared. This is almost what I will do in this section, but I will do so in a way that has more content, by assuming that an economy's technology level is just the average level of its workers' human capital raised to a power. That is, I assume (as in my 1988 paper), that the production function takes the form

$$y = Ax^{\beta}h^{\gamma}, \tag{3}$$

where y is income per effective worker, x is capital per effective worker, and h is human capital per worker. I interpret the term h^{γ} as an external effect (just as in Paul Romer, 1986). It multiplies the productivity of a worker at any skill level h', exactly as does the intercept A in (3).

The marginal productivity of capital formula implied by (3) is

$$r = \beta A^{1/\beta}y^{(\beta-1)/\beta}h^{\gamma/\beta}. \tag{4}$$

I propose to estimate the parameter γ using Edward Denison's (1962) comparison of U.S. productivity in 1909 and 1958, and then to apply this estimate to (4) using Krueger's cross-country estimates of relative human capital stocks in 1959 to obtain a new prediction on relative rates of return on capital.

The estimation of γ is as reported in my earlier paper (1988, p. 23). Using Denison's estimates for the 1909–59 period in the United States, output power man-hour grew about one percentage point faster than capital per man-hour. Denison estimates a growth rate of h, attributed entirely to growth in schooling, of .009. With the technology (3), this implies that $(1-\beta+\gamma)$ times the growth rate .009 of human capital equals .01. With a capital's share $\beta=.25$, these numbers imply $\gamma=.36$. That is to say, a 10 percent increase in the average quality of those with whom I work increases my productivity by 3.6 percent. (This estimate is based on the assumption that the total stock of human capital grows at the same rate, .009, as that part of the stock that is accumulated through formal schooling. I do not have any idea how accurate an assumption this is.)

Now taking the Krueger estimate that five Indians equals one American, the predicted rate of return ratio between India and the United States becomes $(3)^{1.5}5^{-1}=1.04$. That is, taking the external effects of human capital into account in the way I have done entirely eliminates the predicted return differential. Notice that this result is in no way built into my estimation procedure. The value of γ estimated from the 1909–58 U.S. comparison exactly eliminates the return differential in a 1959 India–U.S. comparison.

One might accept this calculation as a resolution of the question I posed in my title. This was the argument in my earlier paper, based on U.S. data only, and I am surprised how well it works in a cross-country comparison. But it is important and troublesome, I think, to note that the cross-country comparison is based on the assumption that the external benefits of a country's stock of human capital accrue *entirely* to producers within that country. Knowledge spillovers across national boundaries are assumed to be zero. Ordinary experience suggests that while some of the external benefits of increases in individual knowledge are local, confined to single cities or even small neighborhoods of cities, others are worldwide in scope. But, without some real evidence on the scope of these external effects, I do not see how to advance this quantitative discussion any further. The argument of this section and the preceding one suggests that correcting for human capital differentials reduces the predicted return ratios between very rich and very poor countries from about 58 at least to about 5, and possibly, if knowledge spillovers are local enough, to unity.

III. Capital Market Imperfections

I have been discussing capital flows in static terms, taking it for granted that differences in marginal products of capital at a point in time imply flows of capital goods through time. In the one-good context I am using, such flows are simply borrowing contracts: the poor country acquires capital from the rich now, in return for promised goods flows in the opposite direction later on.

Suppose countries A and B are engaged in such a transaction, and that the capital stocks in the two countries are growing on paths that will eventually converge to a common value. If we look at goods flows through time between these two countries, we see a phase in which goods flow from advanced A to backward B, followed by a phase (which lasts forever) in which goods flow from B to A in the form of interest payments or repatriated profits. This sort of pattern was implicit in my statement of the capital flow problem. For such a pattern to be a competitive equilibrium, it is evident that there must be an effective mechanism for enforcing international borrowing agreements. Otherwise, country B will gain by terminating its relationship with A at the point where the repayment period begins, and, foreseeing this, country A will never lend in the first place. A capital market imperfection of this type is often summarized by the term "political risk."

A serious difficulty with political risk as an explanation for the inadequacy of capital flows lies in the novelty of the current political arrangements between rich and poor nations. Until around 1945, much of the Third World was subject to European-imposed legal and economic arrangements, and had been so for decades or even centuries. A European lending to a borrower in India or the Dutch East Indies could expect his contract to be enforced with exactly the same effectiveness and by exactly the same means as a contract with a domestic borrower. Even if political risk has been a force limiting capital flows since 1945, why were not ratios of capital to effective labor equalized by capital flows in the two centuries before 1945?

I do not know the answer to this question but, in seeking one, I see no reason to assume that the role of the colonial powers was simply to enforce a laissez-faire trading regime throughout the world. The following monopoly model, very much in the spirit of Adam Smith's (1776/1976) analysis of an earlier phase of colonialism, seems to me suggestive in several ways.

Consider an imperial power whose investors have access to capital at a (first) world return of r. Assume that the imperialist has exclusive control over trade to and from a colony, but that the labor market in the colony is free. Now suppose, at one extreme, that the colony has *no* capital of its own, and no ability to accumulate any. Then capital per worker, x, in the colony can be chosen by the imperialist, and the entire income repatriated. Under these conditions, what value of x is optimal from the viewpoint of the imperial power, viewed as a monopolist?

Let the production function in the colony be $y = f(x)$. Then the monopolist's problem is to choose x so as to maximize

$$(5) \qquad f(x) - [f(x) - xf'(x)] - rx,$$

or total production less wage payments at a competitively determined wage less the opportunity cost of capital. The first-order condition for this problem is

$$(6) \qquad f'(x) = r - xf''(x),$$

so that the marginal product of capital in the colony is equated to the world return r plus the derivative of the colony's real wage rate with respect to capital per worker. It is the imperialist's monopsony power over wages in the colony that is crucial. His optimal policy is to retard capital flows so as to maintain real wages at artificially low levels.

With the Cobb-Douglas technology assumed in my earlier examples, the formula (6) implies that $r = \beta^2 x^{\beta-1} = \beta f'(x)$. With a β value of .4, then, the return on capital in the colony should be about 2.5 times the European return. These are quantitatively interesting rents. The possibility that such rents were important is, I think, reinforced by many of the institutional features of the colonial era: the carving up of the Third World by the European powers, and the frequent granting of exclusive trading rights to monopoly companies.[1]

In a country like India or Indonesia, where most of the workforce was (and still is) engaged in traditional agriculture, it is hard to imagine that the ability to control capital inflows from abroad gave the imperialists much monopsony power over the *general* level of wages. Put another way, the value of capital imported from Europe must have been a small fraction of capital in these countries as a whole, most of which was land. If monopoly control over capital imports was an important source of colonial return differentials, it must have been because only a small part of the colonial labor force was skilled enough to work with imported capital in, say, goods manufacturing. But to explore this possibility, we would obviously need a more refined view of the nature of human capital than one in which five day-laborers equal one engineer.[2]

Insofar as monopoly control over trade in capital goods was an important factor in the determination of capital-labor ratios prior to 1945, I do not see any reason to believe it ceased to be a factor after the political end of the colonial age. Monopoly returns are not of interest to Europeans only: There is much unsystematic evidence of heavy private taxation of capital inflows in Indonesia, the Phillipines, in the Iran of the Shah, and other poor economies that are otherwise attractive to foreign investors. Restrictions on capital flows imposed by the borrowing country are often explained as arising from a mistrust of foreigners or a reluctance to let development proceed "too fast," but I think such explanations warrant a Smithian skeptism.

---

[1] With its emphasis on capital investment, Maurice Dobb's (1945) discussion of late nineteenth and early twentieth-century colonialism is closer to the model in the text than is Smith's. According to Lance Davis and Robert Huttenback (1989), investment in the late British empire was open to firms from any country on competitive terms, which would obviously be inconsistent with this model. Moreover, they do not find rates of return in the British colonies that exceeded European returns for similar investments.

[2] See Nancy Stokey (1988) for a model in which high human capital workers do qualitatively different things than do low human capital workers.

IV. Conclusions

Why does it matter which combination, if any, of the four hypotheses I have advanced is adequate to account for the absence of income equalizing international capital flows? The central idea of virtually all postwar development policies is to stimulate transfers of capital goods from rich to poor countries. Insofar as either of the human capital-based hypotheses reviewed in Sections I and II of this paper is accurate, such transfers will be fully offset by reductions in private foreign investment in the poor country, by increases in that country's investments abroad, or both. Insofar as returns on capital are not equalized, but where return differentials are maintained so as to secure monopoly rents, capital transfers to poor countries will also be fully offset by reductions in private investments. Giving goods to a monopolist does not reduce his interest in exploiting potential rents.

Only insofar as political risk is an important factor in limiting capital flows can we expect transfers of capital to speed the international equalization of factor prices. In a world of largely immobile labor, policies focused on affecting the accumulation of human capital surely have a much larger potential. So too, I think, do policies in which aid of any form is tied to the recipient's openness to foreign investment on competitive terms.

REFERENCES

Davis, Lance E. and Huttenback, Robert A., "Businessmen, the Raj, and the Pattern of Government Expenditures: The British Empire, 1860 to 1912," in David W. Galenson, ed., *Markets in History*, Cambridge: Cambridge University Press, 1989.

Denison, Edward F., *The Sources of Economic Growth in the United States*, New York: Committee for Economic Development, 1962.

Dobb, Maurice, *Political Economy and Capitalism*, Westport: Greenwood, 1945.

Krueger, Anne O., "Factor Endowments and Per Capital Income Differences Among Countries," *Economic Journal*, September 1968, *78*, 641–59.

Lucas, Robert E., Jr., "On the Mechanics of Economic Development," *Journal of Monetary Economics*, January 1988, *22*, 3–32.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–37.

Smith, Adam, *The Wealth of Nations*, Chicago: University of Chicago Press, 1976.

Stokey, Nancy L., "Learning by Doing and the Introduction of New Goods," *Journal of Political Economy*, August 1988, *96*, 701–17.

Summers, Robert and Heston, Alan, "A New Set of International Comparisons of Real Product and Price Levels: Estimates for 130 Countries, 1950–1985," *Review of Income and Wealth*, March 1988, *34*, 1–25.

Are Nonconvexities Important for Understanding Growth?

By PAUL M. ROMER*

When economic reasoning is supplemented only by everyday experience, it leads to few definite conclusions. Systematic data collection and econometric analysis are usually needed to tell whether it is the one hand or the other hand that holds the right conclusion. The title of this note poses one of the few important questions that can be resolved (in the affirmative) using logic and the kind of evidence reported in the newspaper.

I. Are There Nonconvexities?

The *Wall Street Journal* reported last week that two former employees of Du Pont were convicted of extortion. They had threatened to reveal trade secrets about how to make Lycra. Earlier in the year, the paper reported a suit against a former employee of Intel over his alleged use of secrets learned there. It also described an investigation of a firm that used illicitly acquired mechanical drawings and metallurgical formulas to manufacture replacement blades for General Electric turbines.

These incidents have two things in common. First, the value of the economic good at stake is not small. The extortionists asked Du Pont for $10 million. Intel claims that it has invested $100s of millions on the design of its microprocessors. General Electric claims that its investment in the design of turbine blades is worth more than $200 million. Second, the notion of theft is unusual. The companies involved were not deprived of the use of the input that was allegedly stolen. Du Pont could still use its secret process for making Lycra. Intel could still use its chip designs. General Electric could still use its mechanical drawings and metallurgical formulas to make turbine blades.

These examples, and hundreds more like them, show that there are valuable inputs in production that can be used simultaneously in more than one activity. In public finance, consumption goods with this property are referred to as nonrival consumption goods. By analogy, I have referred to goods like a chemical process, a chip design, a mechanical drawing, or a metallurgical formula as nonrival inputs in production (see my 1990 paper).

There are (at least) two ways to think about nonrival inputs. One is to treat a good like a design or a list of instructions as something that is distinct from the medium on which it is stored, and to say that it can be used simultaneously by arbitrarily many different firms and people. A more literal way to describe a nonrival good is to treat the physical medium containing the design or instructions as the relevant good. Then a nonrival input has a high cost of producing the first unit and a zero cost of producing subsequent units. For General Electric, it may have taken millions of dollars of engineering work to produce the first mechanical drawing for its current generation turbine blade, but subsequent drawings can be made at virtually zero cost on a photocopy machine.

Rivalry is closely related to, but distinct from, the notion of excludability. A good is excludable if someone with a property right can exclude others from taking advantage of it. For a rival good, possession ensures excludability. The car I drive is both rival and excludable, since it is relatively easy to possess. The space in which I park my car at a shopping center is also a rival good, but it is one for which it is harder to enforce property rights. At best, it is partially excludable. Just as a good can be rival and nonexcludable, goods can be nonrival and excludable.

---

*University of Chicago and Center for Advanced Study in the Behavioral Sciences, 202 Junipero Serra Blvd., Stanford, CA 94305. Support from NSF grant no. SES-8821943 and BNS#87-00864 and a Sloan Foundation fellowship are gratefully acknowledged, as are the comments of Mancur Olson.

Computer software is a nonrival good, but it can be made excludable using copy protection or a copyright.

Conventional economic goods are rival and excludable. Public goods are nonrival and nonexcludable. Intermediate goods were described by Mancur Olson (1965) and labeled club goods by James Buchanan (1965). (See Todd Sandler and John Tschirhart, 1980, for a review of club theory.) Club goods like the right to swim laps in a pool are perfectly excludable and partially nonrival. Producing the second unit is cheaper than the first, but eventually, either the cost of subsequent swimming rights rises or their quality falls.

For growth theory, it is partial excludability not partial rivalry that is relevant. Goods like new mechanical designs, metallurgical formulas, chemical processes, and semiconductor designs are purely nonrival. They do not wear out or suffer from congestion. In this sense they are like basic science, for example, the advances in solid state physics that led to the discovery of the transistor. The difference is that advances in solid state physics are completely nonexcludable, whereas these goods are at least partially excludable. Otherwise, they would not be intentionally produced by profit-maximizing firms.

The distinction between rivalry and excludability is important because nonrivalry is inextricably linked to nonconvexities, whereas nonexcludability is not. Under the open field system, pasture land is a nonexcludable good that is used inefficiently. Efficiency can be restored by creating a legal system that assigns property rights and establishes excludability. No nonconvexities are involved. No difficulty for price taking arises.

The fact that nonrivalry creates a nonconvexity is apparent from the second definition of a nonrival good as one for which subsequent units have a lower unit cost of production than the first. For the analysis of technology and growth, it is more revealing to consider the definition of a purely nonrival good as one that can be used repeatedly. The most basic premise in our scientific reasoning about the physical world is that it possible to replicate any sequence of events by replicating the relevant initial conditions. (This both a statement of faith and a definition of relevant initial conditions.) For production theory, this means that it is possible to double the output of any production process by doubling all of the rival inputs. By definition, the nonrival inputs can be used again in the replica process. Hence, if there are any nonrival inputs that have productive value, output will increase more than proportionately with an increase in all of the inputs. If R denotes the set of rival inputs, N the set of nonrival inputs, and $F(R,N)$ denotes output, then for integer values of λ,

$$F(\lambda R,\lambda N) > F(\lambda R, N) = \lambda F(R,N).$$

This means that in the large, the elasticity of output with respect to inputs is greater than 1 and that the function $F(\cdot)$ is not concave.

The structure of this argument is very simple. Once you define the notion of a nonrival good and admit the principle of replication, it follows that nonconvexities are present. Newspaper reports remind us that nonrival goods exist and are important.

II. Previous Statements of This Argument

So far as I know, the statement of this argument in terms of nonrivalry is new, but the basic idea is old. Joseph Schumpeter suggested something like this in 1942. Richard Nelson and Sidney Winter have long recognized its importance for growth and industrial organization (1982). William Nordhaus (1969) noted this difficulty in his discussion of invention, and Karl Shell emphasized it in his models of growth (1966, 1967, 1973). Mancur Olson (1973) discussed the nonconvexity associated with public goods, as did David Starrett (1977).

In an influential discussion of these issues, Kenneth Arrow (1962) used "appropriability" in place of "excludability" and "indivisibility" in place of "nonrivalry." Appropriability is synonymous with excludability, but indivisibility is not quite precise enough to capture what is really different about goods like designs. Indivisibilities typically imply local nonconvexities. Nonrival goods create global nonconvexities.

Perhaps because the notion of rivalry remained obscure, or perhaps because property rights and excludability are easy to contemplate without undermining the notion of a price-taking equilibrium, what has come to be known as the "appropriability framework" (David Mowery and Nathan Rosenberg, 1989) has focused on excludability, and has tended to neglect the essential nonconvexities noted clearly by Arrow and others.

III. What About Fixed Factors and Aggregation?

The claim that nonconvexities are unavoidable has not gone unnoticed by defenders of price taking. At least three lines of attack on the argument outlined here can be discerned. They are: 1) that there are fixed factors of production; 2) that nonrival inputs are hard to measure; and 3) that the nonconvexities may be present, but they are really not very important for aggregate level analysis.

Because there are factors of production that are in fixed aggregate supply, it is certainly true that it is not possible to replicate all productive activities at the same time. This is a red herring. Nothing in the logic of the argument establishing the nonconvexities requires that it actually be possible to double all inputs. What matters for an equilibrium is what happens to output in any particular production activity if its inputs are doubled, for example, by taking inputs away from some other activity. In any activity, Euler's theorem implies that an elasticity of output with respect to all inputs that is greater than one makes it infeasible to pay each input a rental rate equal to its marginal productivity. Factors like land or managerial ability that are asserted to be in fixed supply have alternative uses in production (and consumption), and can be shifted between firms engaged in the same activity. How they are allocated depends on how they are compensated, and this depends crucially on convexity.

It is also true that the aggregate quantity of nonrival inputs like processes, formulas, instructions, and designs is hard to measure. But, in the years since the capital controversies, the assertion that difficulties of measurement and aggregation can be used to attack theoretical propositions has worn rather thin. These difficulties are hardly unique to nonrival goods. If you look carefully and worry about things like quality change and the introduction of entirely new goods, even concepts like GNP are plagued with ambiguity.

Concepts that are vague at the aggregate level are obvious at the industry or firm level. For the personal computer industry, it is clear what we mean by changes in output or inputs. Industry output (measured in millions of instructions per second of capacity produced per year) can double because rival inputs double; twice as much aluminum, silicon, copper, steel, and plastic can be used along with twice as many workers and machines to produce twice as many physical PCs. Or nonrival inputs can go up when Intel produces a new design for a microprocessor that lets each new PC run twice as fast.

IV. But Do the Nonconvexities Really Matter?

The fallback position to which some defenders of price taking retreat is this: Yes, we admit that nonrival inputs like new chemical reactions, new electronic devices, and advances in materials science will revolutionize living standards within each individual's lifetime. And yes, these inputs are intrinsically tied to nonconvexities. But no, we do not think that this poses a problem for aggregate equilibrium theories based on price taking.

The argument that nonconvexities might not matter at the aggregate level seems to be based on two inappropriate analogies. The first is a confusion of nonrivalry with excludability. If excludability were the only problem, policy towards research and development would be just like policy toward open fields and national defense: Establish property rights if this is possible; have the government pay for a good if it is not. If the problem was merely one of excludability, price taking would be observed in the markets for all of the excludable goods. Optimal provision of the public good by the government would then lead to a Pareto optimal equilibrium.

The deeper problem here is nonrivalry and its associated nonconvexity. Making a nonrival good excludable does not change the fact that it is nonrival. For a firm like Intel, selling its goods at marginal cost is not a viable strategy. Giving Intel better copyright or patent protection may be a good idea, but it will not cause it to behave like a price taker.

A related argument is that departures from the perfect market assumptions must be small if there is no evidence of spillovers or external effects. What the evidence is on spillovers is not entirely clear. In any case, the logic of the argument is wrong. An economy with perfect patent protection for ideas must exhibit departures from price taking. Patents are supposed to create monopolies. Even if patent protection is perfect and there are no external effects and no spillovers, the associated equilibria will generally not be Pareto optimal. (For an example of an economy of this type, see my 1987 paper.)

Economists who recognize the presence of nonconvexities sometimes rely on a second analogy to argue that they are not important at the aggregate level. The canonical example of a nonconvexity that can be convexified with large numbers is an indivisibility that gives firms a U-shaped average cost curve. As the number of identical firms become very large, price taking can be supported at a price equal to minimum average cost. This kind of analysis has been used in club theory to show that private clubs can provide the optimal level of an excludable, partially nonrival good.

The point that should be clear is that not all nonconvexities can be convexified this way. Large numbers of potential firms do not help if the average cost curve of each individual firm is globally decreasing. This is exactly the case that arises with nonrival goods. To see the problem from the production side, let $\bar{x}$ be the amount of an input x needed to create a nonrival good y. Then the production possibility set is the union of the sets $\{(x,0)\colon\ x \in [0,\bar{x})\}$ and $\{(x,y)\colon\ x \geq \bar{x}$ and $y \geq 0\}$ in $\mathbb{R}^2$. The closed convex hull of this set is the nonnegative orthant in $\mathbb{R}^2$. If the original production set is convexified, the implied technology says that any amount of the good y can be had for free. The only supporting price for y in terms of x is zero, and a zero price is inconsistent with an equilibrium in which firms earn nonnegative profits.

V. What Is the Alternative to Equilibrium Theory with Price Taking?

It is clear that we can do general equilibrium theory without assuming price taking. The relevant point was emphasized by Joseph Ostroy (1973, 1984): If an economy has many goods as well as many agents, markets for some goods may still be thin, even in the limit as the number of goods and agents goes to infinity. Market power can survive. But in the limit economy, individual agents are small relative to the economy as a whole, so these "large-square" economies avoid many of the ambiguities that arise in strategic situations with small numbers of agents.

It is not hard to construct large-square growth models that are directly comparable with conventional growth models. Examples are given in my 1987 and 1990 papers, by Philippe Aghion and Peter Howitt (1989), and by Gene Grossman and Elhanan Helpman (1989c). These models differ in the assumptions about spillovers (the last three have them, the first does not) and about the form of innovations (entirely new goods in my models; replacements for old goods in the other two). A product life cycle model that is closely related to the growth models cited here was described by Paul Segerstrom et al. (1987).

All of the models have nonrival goods, so they will share a result about trade and growth. Suppose that two identical economies operate in total isolation from each other. Because they are identical, they allocate the same resources to the research sector and make the same discoveries in the same order. In the classical theory of trade with a convex technology, there are no gains from trade between identical economies. Here, there are large potential gains from interaction. Because there are no limits on the use of nonrival goods, there is no reason to have engineers in the different countries solve the same problem twice. It would be feasible for the engineers of one country to supply the whole world with the same level of nonrival goods as was produced using both sets of engineers under total isolation. The worldwide equilibrium with isolation must be Pareto suboptimal because the engineers in the other country could go on vacation without in any way reducing worldwide output. If increasing the inputs into research is good for growth and welfare, the second set of engineers could be set to work on different problems, thereby doubling the effective resources allocated to research.

The fact that isolation is suboptimal does not necessarily imply that opening trade will be welfare improving. Even if it is, free trade will not lead to an equilibrium that is first-best optimal because of the nonconvexities and departures from price taking that are present. Nonetheless, what seems to be a robust result is that trade in goods between similar countries will lead to a welfare-improving reallocation of resources used in research (see my paper 1990; Luis Rivera-Batiz and myself, 1989). Because the countries are similar, the main effect of opening them to trade will be to induce trade in goods developed through research, thereby reducing redundancy in research effort.

When the countries are different (for example, when one has a higher ratio of skilled human capital to raw labor) opening trade will also affect the relative price of human capital and labor. This can have an ambiguous effect on the market incentives for undertaking research. If an abundant supply of cheap labor reduces the incentive to produce nonrival inputs (as in the models described in my 1989 paper and in Grossman-Helpman, 1989c), then for a country like the United States, increased trade with Mexico might slow down the rate of growth even if increased trade with Europe or Canada would speed it up.

Without extensive (and implausible) government intervention, equilibria in this kind of economy will not be first-best Pareto optimal. As a result, economic reasoning alone can lead to a great variety of counterintuitive second-best results. The rate of growth and total research effort might be too high even if there are positive knowledge spillovers (Aghion-Howitt and Grossman-Helpman, 1989c). Tariffs on some of the goods produced in a country might speed up growth by driving human capital out of production of these goods and into research (Grossman-Helpman, 1989b). Allowing for reverse engineering and knockoffs of goods by less developed countries might speed up the rate of introduction of new goods in the developed country; even though the return to introducing a new good will be lower, the cost of inputs into research might be lower. Labor that is forced out of manufacturing of goods by the foreign goods will drive down wages, and some of this labor may shift into research (Grossman-Helpman, 1989a).

Whether these theoretical possibilities actually occur is the kind of question that does require systematic data collection and econometric analysis. I have argued (1989) that there is evidence for a negative effect of labor growth on technological change. Across countries, across sectors in the United States, and over time in the United States, higher rates of labor force growth are associated with lower rates of productivity growth. There is a large body of work that tries to measure the rate of return to research (Zvi Griliches, 1988). Both the private and the social rate of return to research seem high relative to other rates of return, so there is a presumption that too few rather then too many resources are devoted to research. This is at least a case in which the formal evidence jibes with the newspaper reports. As a fraction of GDP, Japan now spends considerably more on commercial (not basic) research that generates excludable benefits than does the United States, and the sentiment on both sides of the Pacific seems to be that they are better off as a result.

The other conjectures remain open. Whether increased trade generally increases growth rates is still a matter of some debate. Whether copying and product cycle effects speed up or slow down growth has to my knowledge not been studied in detail.

VI. Conclusion

The oldest question in economics is what causes growth. One of the oldest conjectures, built into Adam Smith's story of the pin factory, is that nonconvexities are important for growth and that they create an incentive for international trade that increases the extent of the market. (For a recent reminder of the fact that nonconvexities are essential in the pin factory story, see the discussion by Brian Edwards and Ross Starr, 1987.) We now know how to fit this kind of effect into an aggregate growth model, and we can already see that these models generate many theoretical possibilities. Whatever the verdict on the details, we have enough evidence in hand to come to a consensus about whether nonconvexities matter for growth. It seems clear that nonrival goods exist, that they are important for aggregate growth, and that they create nonconvexities that matter for aggregate level analysis. Moreover, there is some reason to believe that Smith's conjecture about the extent of the market is right. The very notion of nonrivalry suggests that there are large dynamic gains from trade between similar countries.

REFERENCES

Aghion, Philippe and Howitt, Peter, "A Model of Growth Through Creative Destruction," MIT working paper, 1989.

Arrow, Kenneth, "Economic Welfare and the Allocation of Resources for Invention," in Richard Nelson, ed., *The Rate and Direction of Inventive Activity*, Princeton: Princeton University Press, 1962.

Buchanan, James A., "An Economic Theory of Clubs," *Economica*, February 1965, *32*, 1–14.

Edwards, Brian K. and Starr, Ross M., "A Note on Indivisibilities, Specialization, and Economies of Scale," *American Economic Review*, March 1987, *77*, 192–95.

Griliches, Zvi, *Technology, Education and Productivity*, New York: Basil Blackwell, 1988.

Grossman, Gene and Helpman, Elhanan, (1989a) "Endogenous Product Cycles," NBER Working Paper No. 2913, 1989.

______ **and** ______, (1989b) "Growth and Welfare in a Small Open Economy," NBER Working Paper No. 2970, 1989.

______ **and** ______, (1989c) "Quality Ladders in the Theory of Growth," NBER Working Paper No. 3099, 1989.

Mowery, David C. and Rosenberg, Nathan, *Technology and the Pursuit of Economic Growth*, Cambridge: Cambridge University Press, 1989.

Nelson, Richard R. and Winter, Sidney G., *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

Nordhaus, William, *Invention, Growth, and Welfare*. Cambridge: MIT Press, 1969.

Ostroy, Joseph, "Representations of Large Economies: The Equivalence Theorem," presented at the winter meetings of the Econometric Society, 1973.

______, "A Reformulation of the Marginal Productivity Theory of Distribution," *Econometrica*, May 1984, *52*, 599–630.

Olson, Mancur, *The Logic of Collective Action: Public Goods and the Theory of Collective Action*, Cambridge: Harvard University Press, 1965.

______, "The Measurement of Economic and Social Performance," in Milton Moss, ed., in *Evaluating Performance in the Public Sector*, New York: Columbia University Press, 1973.

Rivera-Batiz, Luis and Romer, Paul M., "International Trade with Endogenous Technological Change," University of Chicago working paper, 1989.

Romer, Paul M., "Growth Based on Increasing Returns Due to Specialization," *American Economic Review Proceedings*, May 1987, *77*, 56–62.

______, "Endogenous Technological Change," *Journal of Political Economy*, forthcoming 1990.

______, "Capital, Labor, and Productivity," prepared for the December meeting of the Brookings Microeconomics Panel, 1989.

Sandler, Todd and Tschirhart, John, "The Economic Theory of Clubs: An Evaluative Survey," *Journal of Economic Literature*, December 1980, *18*, 1481–521.

Schumpeter, Joseph, *Capitalism, Socialism, and Democracy*, New York: Harper, 1942.

Segerstrom, Paul S., Anant, T. C. A. and Dinopoulos, Elias, "A Schumpeterian Model of the Product Life Cycle," Econometrics and Economic Theory Paper no. 8606, Michigan State University, 1987.

Shell, Karl, "Toward a Theory of Inventive Activity and Capital Accumulation," *American Economic Review Proceedings*, May 1966, *56*, 62–68.

______, "A Model of Inventive Activity and Capital Accumulation," in his *Essays in the Theory of Optimal Economic Growth*, Cambridge: MIT Press, 1967.

______, "Inventive Activity, Industrial Organization, and Economic Growth," in J. Mirrlees and N. Stern, eds., *Models of Economic Growth*, New York: Wiley & Sons, 1973.

Starrett, David A., "Measuring Returns to Scale in the Aggregate and the Scale Effect of Public Goods," *Econometrica*, September 1977, *45*, 1439–55.

LESSONS FOR DEVELOPMENT FROM THE EXPERIENCE IN ASIA[†]

Lessons From Korean Economic Growth

By SUSAN M. COLLINS*

In 1960, Korea was a poor developing country with a small manufacturing sector and heavily dependent on foreign aid. It had seemingly few prospects to increase and maintain high growth rates. However, between 1965 and 1979, Korea's real GDP growth averaged over 9 percent per year, with manufacturing growth of nearly 19 percent.

In 1981, Korea was the fourth largest debtor country in the world, behind Brazil, Mexico, and Argentina. Output had declined sharply in 1980. With a debt-to-GDP ratio of 50 percent, there were widespread concerns about Korea's ability to meet its debt obligations. However, Korea's economy was booming again by 1986, with a substantial trade surplus due to rapid export growth. In addition to meeting all debt service obligations, Korea had begun to repay the principal on its external debt.

Korea's impressive history of economic growth stands out from the experience of most developing countries that have borrowed heavily in international financial markets. A comparison of per capita income (PCI) growth rates from the World Bank presents the differences starkly. Over 1965–87, Korea's PCI grew at an average annual rate of 6.4 percent. Over the same period, PCI grew on average by just 2.0 percent in the 17 heavily indebted countries, by 2.5 percent in middle-income countries and by 3.1 percent in low-income countries. (In fact, only three World Bank member countries—Botswana, Oman, and Singapore—had faster annual PCI growth rates than Korea.)

As stated in the *World Development Report*, "for most of the highly indebted countries, the debt crisis has become a growth crisis as well" (1989, p. 17). The contrasts in growth performance between these countries and Korea, since 1965 and particularly during the 1980s debt crisis recovery, are striking. This paper examines distinguishing aspects of Korea's economic growth and argues that these factors do contain lessons for policymakers in other countries struggling to revive stagnant economies.

I. Policy Stability

Korea's experience does not provide an example of "quick fix" policy packages, but of a long history of relatively consistent, stable, and sensible macroeconomic policies. At the same time, Korea's history is not completely devoid of policy mistakes. For example, the 1973–79 "Big Push" to develop heavy industry is widely recognized as having contributed to a real appreciation and loss of competitiveness and to distorting allocation of credit and other resources.[1] Instead, what stands out is that macroeconomic policies were adjusted before they became far out of line. The point could be made with a variety of policy indicators; budget deficits, real interest rates and real exchange rates are discussed below.[2]

First, budget deficits have been kept small in Korea, averaging just 2.3 percent, and

---

[†] *Discussant*: Andrei Shleifer, University of Chicago.

*Harvard University, Cambridge, MA 02138.

[1] See my paper with Won-Am Park (1989, especially pp. 191–98) for further discussion of Korea's Big Push and also for additional references.

[2] Data on central budget deficits and real interest rates come from the International Monetary Fund. Real exchange rate data are from Morgan Guaranty.

ranging from 1.0 to 4.2. The average deficit was more than twice as large in Mexico and Argentina. The range between the smallest and the largest deficit was more than 10 percent of GNP in all three Latin American countries.

Korea has also maintained relatively stable real interest rates. Between 1977 and 1987, the real deposit rate (average nominal rate less *ex post* consumer price inflation) averaged 2.5 percent, compared to −2.3 percent in the Philippines, −13.4 percent in Mexico, and −22.3 percent in Argentina. While Korea's real deposit rates were not positive every year, they were negative in only two of eleven years, compared with five years of negative rates in the Philippines, six years in Argentina, and nine years in Mexico.

Relative to other countries, Korea's real exchange rate has remained stable as well. Although a fixed nominal exchange rate was maintained from December 1974 through January 1980, resulting in a 15 percent real appreciation, this appreciation is small when compared to the real appreciations of 23 percent in Mexico during 1977–81, and the 81 percent in Argentina during 1977–80. The standard deviation in Korea's real exchange rate during 1975–85 was just 6.5, less than half of the 14.7 and 16.7 standard deviations in Brazil and Argentina.

II. Breathing Space for Policy Reforms

In 1980, Korea was in the midst of an economic crisis. Real output declined by 4.8 percent and inflation had doubled to 28.7 percent from 14.4 percent in 1978. The debt/GNP ratio had jumped from 28 to 45 percent and the current account deficit had ballooned to 9 percent of GNP. Years of the Big Push to heavy industry had resulted in growing price controls, import restrictions, and regulations distorting financial markets. Korea needed a combination of macroeconomic stabilization policies, trade and financial market liberalization, and economic restructuring. Korea's recovery from this crisis was striking and rapid. Real growth averaged over 7 percent during 1981–84 and, by 1984, inflation had been cut to 2.3 percent, the current account deficit was less than 2 percent of GNP, and there was little remaining concern about Korea's creditworthiness.

Should we believe that other countries could implement the same policy package to rapidly restore creditworthiness and growth? My response has three parts. First, Korea's policy package is a sound one for other countries. However, Korea's history of relatively stable, sensible macroeconomic policies, high rates of investment, and strong growth also played a role in the rapid turnaround. In this sense, Korea's short-lived economic crisis was less severe than the difficulties faced by many heavily indebted countries. The second point, therefore, is that Korea's policies are unlikely to result in as quick and as impressive a recovery in countries with a history of policy reversals and inconsistencies, and which have had low investment rates since 1982.

The third point is that many observers omit a key part of Korea's recovery package. Korea did not implement macroeconomic stabilization and restructuring measures, nor did it revive growth at the same time that it reversed its dependence on foreign borrowing. Instead, Korea had the breathing space of large continued capital inflows (together with some special circumstances) during the first years of its recovery which enabled it to revive growth *before* undertaking restrictive monetary and fiscal policy actions. Korea does *not* provide an example of how to stabilize and restructure an economy, shift from receiving capital inflows to making large resource transfers abroad and raise living standards, all at the same time.

Table 1 provides economic indicators for Korea during 1979–84. It is useful to divide the recovery into two periods: 1980–82 and 1983–84. In the early period, Korean policymakers devalued, liberalized price controls and many import restrictions, and initiated tax reforms. Although monetary and fiscal policies were initially to be restrictive, the tight stance was relaxed because of concerns over flagging investment in 1981, and to bail firms out after a financial crisis in 1982. In fact, fiscal policy was quite expansionary,

TABLE 1—KOREAN ECONOMIC INDICATORS

	1979	1980	1981–82	1983–84
GNP Growth Rate	7.0	−4.8	6.0	10.2
Current Account	−6.8	−8.8	−5.4	−1.9
Fixed Investment	33.2	32.3	29.6	31.3
Domestic Saving	28.1	23.5	23.8	29.1
*M*2 Growth Rate	24.6	26.9	26.0	11.5
Budget Deficit	−1.4	−3.2	−4.6	−1.5
External Debt	32.5	45.0	53.5	52.3
External Debt (billions $)	$20.3	$27.2	$37.1	$43.1

Source: Economic Planning Board, Korea.
Notes: Percent of GNP, unless otherwise indicated; 1981–82 and 1983–84 data are annual averages, except for external debt, which is end of period.

accommodated by strong money growth. Korea continued to borrow heavily to finance imports and investment. Also, a good harvest (following extremely poor harvests in 1978 and 1980) helped to expand output while reducing food imports.

The 1981–82 measures set the Korean economy up to take advantage of stronger world growth in 1983–84. It was not until exports and output were booming that the government reversed its expansionary macroeconomic policies, and the debt-to-GNP ratio stabilized.

Korea's phased policy response, with its initial breathing period, contrasts sharply with the attempts in many heavily indebted countries (for example, in Latin America) to do everything at once. In these countries, trade deficits were quickly transformed to surpluses after 1982 through cuts in imports and investment. This approach has not produced adjustment with growth.

Korea is not the only example that breathing space facilitates a growth recovery from a debt crisis. Turkey, during 1980–82, and Indonesia, during the mid-1960s, also received generous capital inflows as they were implementing their adjustment packages. Like Korea, these countries both got into trouble *before* the 1982 widespread debt crisis, and both benefited from favorable borrowing conditions in international financial markets. However, breathing space is not a panacea. There are also examples of countries that received large capital inflows, but did not implement necessary reforms, and therefore did not engineer a recovery to sustainable growth and creditworthiness.

III. Investment and Growth

A related feature that distinguishes Korea's adjustment from the experience in many Latin American countries is that Korea maintained high rates of investment. This investment, concentrated in export industries, helps to account for Korea's rapid growth during the 1960s and 1970s, and also for the quick recovery in the 1980s. Korea's gross domestic investment grew at an average annual rate of 15.9 percent during 1965–80 and 10.0 percent during 1980–1987. In contrast, investment growth in the 17 highly indebted countries fell from 8.6 percent during 1965–80 to −5.1 percent during 1980–87. Many analysts have expressed grave concern about their ability to revive growth as the capital stock erodes. External borrowing helped to finance Korea's investments in 1980–82. It was not until the 1983 surge in growth that domestic savings began to rebound, reducing the need for foreign financing. Breathing space contributed significantly to Korea's ability to maintain high rates of investment.

What are the underlying sources of Korea's growth, and how important was capital accumulation? A Denison growth accounting analysis by K. S. Kim and J. K. Park (1985) decomposes Korean growth during 1963–82. They show that factor inputs, and especially capital accumulation, played an increasingly central role. During 1963–72, increased quantity and quality of capital and labor accounted for about half of average annual growth, or 4.2 percent per year. During 1972–82, factor inputs accounted for nearly 80 percent of Korean growth, or 5.6 percent per year. In particular, increased nonresidential structures and equipment augmented growth by 1.5 percent per year more during 1963–72 than during 1972–82, offsetting a decline in factor productivity growth that followed the first oil price shock. On the labor input side, increased work hours and improved education helped to offset slower employment growth. The importance of physical and especially human capital accumulation distinguish Korea's growth history from that of countries which industrialized prior to 1973.

IV. Concluding Remarks

What, then, are the lessons from Korea's economic growth? First, a stable policy environment provides a solid stage for adjusting to internal and external shocks. Although difficult to quantify, there can be little doubt that Korea's policy history was an important stabilizing factor in the quick 1981–84 recovery. Second, investment in both physical and human capital is a key to economic growth, In Korea, increased factor inputs alone accounted for an average annual growth rate of over 5 percent during the 1970s.

It is against this backdrop that a third, important lesson emerges. Even in Korea, where policies have been stable and high rates of investment have been maintained, breathing space in the form of continued capital inflows played a key role in reviving and sustaining growth after 1980. Korea's experience should not be construed as an example that a country in the midst of a prolonged economic crisis, with a depleted capital stock and a history of policy reversals and mistakes, can simultaneously undertake structural adjustments together with restrictive macroeconomic policies, transfer resources abroad and revive stagnant growth rates.

REFERENCES

Collins, Susan M. and Park, Won-Am, "External Debt and Macroeconomic Performance in South Korea," in Jeffrey D. Sachs and Susan M. Collins, eds., *Developing Country Debt and Economic Performance,* Vol. 3, Chicago: University of Chicago Press, 1989.

Kim, K. S. and Park, J. K., *Sources of Economic Growth in Korea: 1963–82*, Seoul: Korea Development Institute, 1985.

Economic Planning Board, *Major Statistics of Korean Economy*, Seoul, Korea, 1988.

World Bank, *World Development Report*, Washington: World Bank, 1989.

Asian Trade and Growth Lessons

By ANNE O. KRUEGER*

The Asian experience with trade and growth has been highly varied. There is a strong contrast between the success of the East Asian "superexporters," the still-impressive but less rapid growth of Southeast Asia, and the sluggish growth of much of the South Asian subcontinent, including especially India, Burma, Nepal, and Bangladesh.

No observer of the success of the East Asian exporters can doubt that exports played a key role in accelerating growth. Although the relationship between growth of real per capita GNP and of exports is certainly not perfect, Korea's export growth rate of 22.9 percent accompanied an average annual growth rate of per capita income of 6.4 percent over the 1965–87 period. Hong Kong, Singapore, and Taiwan were also rapidly growing economies whose growth of exports outpaced that of income. By contrast, the Philippines, Sri Lanka, and Pakistan all experienced annual rates of growth of per capita income of less than 3 percent, and had falling shares of exports in GNP. For Bangladesh and Nepal, growth rate and export performance were even worse.

It bears pointing out that the successful East Asian countries had their earlier experience with inner-oriented trade strategies, quantitative controls over imports, and slow growth. In each case, the shift in trade strategy was accompanied by a sizeable increase in the rate of growth of output, in the marginal productivity of capital as judged by the behavior of the incremental capital output ratio, and in total factor productivity.[1]

The interesting questions and lessons therefore arise with respect to how export growth was achieved, and the reasons why export growth was associated with rapid growth of GNP. Consider first the determinants of rapid export growth. Government policies in all the superexporters were clearly set within the context of a commitment to growth through exporting. Several policy instruments were geared to those targets: the real exchange rate received by exporters, the import regime as it affected reexports, and the infrastructure (ports, domestic transport, communications with the outside world by phone and mail) necessary for exporting all were determined by the needs of the export drive. (See Kwang Suk Kim, 1987, and Richard Snape, 1988, for documentation.)

In every country with a successful development-through-exporting strategy, exporters were exempted from whatever restrictions prevailed in the import regime. This was not only necessary for success in exporting, but it restricted the extent to which officials could monitor and regulate imports. Further, the government's commitment to exporting was an implied assurance of the profitability for successful exporting firms. While some of the NICs initially provided export subsidies, subsidized credit, and other special privileges for exporters, all of these became costly as the export drive succeeded, and there was therefore perpetual pressure on government officials to increased reliance on the exchange rates and uniformity of incentives. In Korea, for example, Kim (Table 2.10) estimates that the effective exchange rate for exports was within 6 percent of that for imports for every year after 1965; the variation of incentives between import-competing and exportable commodities was far smaller than in the typical import-substitution regime.

It is certainly not correct to state that the governments in the export-strategy countries

---

*Duke University, Durham, NC 27706.

[1] For a more systematic account of performance than is permitted by space limitations here, see Paul Kuznets (1988).

provided more incentives for exporting than for import-competing production. What appears to have been the case is that incentives were biased very little away from exporting, and that whatever profits were to be made through import restrictions (in the early stages of the exporting strategy) were channeled to exporters. In the jargon of international trade theory, domestic incentives were reasonably uniform.

It is equally incorrect to state that governments decided which commodities should be encouraged for exporting in the early years of the export-oriented drive. Export subsidies, credit subsidies, and tax exemptions applied to all exporters. Moreover, given the small base on which exports began, any producer who could export was encouraged to do so. (See Jung-ho Yoo, 1989.)

Finally, it must be recognized that in the early years of export-oriented strategies, all relevant participants in the then-poor countries (government officials and producers) perceived themselves to be weak and ineffective in the international market. This perception was significant as a backdrop against which economic policy was framed. Moreover, it permeated the entire structure of governmental decision making, another factor contributing to overall economic growth.

The factors enumerated thus far were all similarities among the successful exporting countries. It may be noted that there were also differences. Chief among these was the contrast in firm size between Korea and Singapore on one hand, and Taiwan and Hong Kong on the other. There were also major differences in the treatment of foreign investors.

It may be concluded that there were a number of factors that contributed to the rapid growth of exports. Certainly, that the initial situation was one in which incentives were highly distorted was one factor. This permitted large efficiency gains. Once rapid growth of exports and real GNP had begun, there was widespread acceptance of the economic strategy which enhanced its credibility, and intensified the mandate of government officials to maintain the momentum of the drive. In all of this, incentives were fairly uniform; decisions were taken regarding prospects for *exporting*; not for particular industries. In addition, labor was relatively abundant, the labor market functioned reasonably well, and hence numerous profitable export opportunities arose.

Interest in the question of why export growth was so rapid has been motivated largely by interest in how other governments' trade policies might be successfully formulated (and in particular the extent to which "picking the winners" is a key element). There has been significantly less effort made to understand how export growth led to such spectacular rates of overall economic growth.

At a superficial level, there are several obvious, and trivially correct, partial explanations for the association between export growth and GNP growth: 1) the East Asian economies had highly protective trade regimes before they embarked on their export-oriented trade and development strategies, and by shifting strategies they achieved gains through realizing their comparative advantage, à la trade theory; 2) associated with that, the incremental capital-output ratio fell sharply, thereby assuring a rapidly rising growth rate even at a constant savings rate; 3) the reliance upon an outer-oriented trade strategy permitted the realization of economies of scale, and the exhaustion of indivisibilities, that would otherwise have arisen; and 4) avoidance of "stop-go" policies associated with balance of payments difficulties that have plagued other developing countries also contributed to more rapid growth.

That these factors were highly significant, especially in the first decade after adopting outer-oriented growth strategies, cannot be doubted. However, questions remain: very rapid growth rates of both exports and real GNP have been sustained for several decades; many of the gains attributed to the four explanations were surely largely exhausted after two decades of rapid growth.

One of the reasons it is difficult to identify the links between export growth and overall growth is that other policies adopted in the superexporting countries have also been conducive to growth. Not only did the trade and payments regime provide reasonably uniform incentives for most tradables, but the labor market worked fairly efficiently, as already mentioned. In addition, observers have pointed to a host of other factors: educational attainments were quite high; governmental services, including transport and communications, were reasonably efficient; even reactions to changes in the external environment, such as the oil price increases of 1973 and 1979, evoked far swifter and more effective adjustments than in many other developing countries.

Moreover, even in areas of policy (such as financial markets and credit rationing) where initial regulations and controls were not conducive to economic efficiency, there has been a continuing process of liberalization as growth has proceeded. In Korea, real interest rates were negative until 1964, and were positive thereafter; credit rationing thus had less scope for misallocating capital after 1964 than it did earlier. By the early 1980s, financial deregulation was proceeding fairly rapidly, and continues even now. Concurrently, there has been a clear tendency to liberalize international financial transactions, to liberalize the trade regime, and to reduce the domain within which direct controls affect economic activity. Similarly in Taiwan, financial liberalization, tax reform geared to reducing the incentives for misallocation of resources, continuing liberalization of imports and capital account transactions, were hallmarks of the continuing rapid rate of growth. Even the rate of inflation has approached very close to zero in the past several years.

Although there are many questions about these policies and their effects, it is possible to put forward a hypothesis about trade, economic growth, and government policies that is consistent with the trends in the superexporters. It has two parts: first, one of the reasons for the success of the countries adopting outer-oriented trade strategies is that an export orientation imposes a discipline and set of constraints on all economic policies that prevent the adoption of very many measures severely antithetical to growth; and second, the extent of liberalization and smooth functioning of markets that is consistent with rapid economic growth increases with the level of output per capita.

Both of these hypotheses are speculative, the second more so than the first. If they are correct, they would suggest a new value of an open economy: not only does openness increase competition, permit the realization of comparative advantage, and provide a means for realizing economies of scale and exhausting indivisibilities, it also constrains government policies in ways that make them conform more closely to those conducive to increases in living standards and rapid economic growth.

Turn first to the ways in which a commitment to an outer-oriented trade strategy imposes a constraint on policies. The fact that anything more than moderate restrictions on imports is inconsistent with such a strategy has already been mentioned. Equally important, however, is the fact that the costs of mistakes become visible fairly rapidly. The mistaken Korean policy of encouraging heavy and chemical industry development in the late 1970s is an example. It was the openness of the economy that made the costs of the policy highly visible (both to policymakers and to most informed observers of economic activity) quickly. Likewise, when the Singaporean authorities decided to adopt a "high wage" policy in the mid-1980s, growth slackened rapidly, and the policy was reversed. Pressures for realistic exchange rates, budgetary policies, and for restraint in attempting to decide which industries should be developed were strong in all the successful exporting countries because of the fact that they were highly open. Given the huge mistakes in policy that seem to have enormous costs in terms of foregone output and growth in many other developing countries, the fact that an outer-orientation constrains policymakers from making these same mistakes, at least to the same degree, is doubtless one of the reasons for more rapid overall growth.

That heavy-handed direct controls are more inconsistent with an export-oriented trade and growth strategy than with an inner-oriented trade strategy may be tautological. The question really centers on the extent to which it is "better policy," or at least "less bad" policy, that brings about the very high rates of growth, and what the causation is. It could be argued that governments that are able to inaugurate and effect export-oriented trade policies sufficiently to induce any lasting success will also be able to carry out other policies. All that can be said with confidence is that the political penalties for adopting inappropriate policies will be greater in the context of an ongoing outward-oriented strategy. In that sense, success may breed success.

The second hypothesis (that the same policies have different economic costs at different stages of development) is the more tentative of the two. Here, the argument would be that at very low-income levels, most economic activity takes place in agriculture. The costs of direct controls over trade, industry, and factor markets are therefore probably fairly small. Until mechanisms are found that permit satisfactory growth of agricultural output and productivity (or leave the agricultural sector relatively backward and inaugurate industrial growth), there can be very little economic growth. In most countries, this means that a prerequisite for satisfactory economic growth will be the provision of infrastructure (including especially education and health, as well as roads, agricultural research and extension, etc.) and incentives for agriculture.

Once that is accomplished, a given set of restrictions over trade will be consistent with an evermore slowly growing set of industrial activities. For, to the extent that growth occurs, a by-product is the increasing complexity of the industrial production structure, and thus the increasing inefficiency of trade restrictions and other given interventions. A move to an export-oriented trade strategy removes that set of high-cost impediments to growth; at first, the move to an outer-oriented strategy can result in large gains in exports even when credit rationing, duties on imports, and restrictions on capital flows are still in force. Over time, however, as the range of industrial activities (and the service activities supporting them) increases, the costs of these restrictions in turn mount. Sustaining growth therefore entails further liberalizations in the import regime, in financial and capital markets, and in controls in international capital flows as the complexity of the domestic economy increases.

For Hong Kong, there has been virtually complete laissez-faire throughout the thirty-year period of rapid development. For Korea, Singapore, and Taiwan, however, there has been continuing liberalization of economic activity. Certainly this liberalization increased the returns to the trade policies that resulted in a balance in incentives for production of import-competing and exportable goods. Whether the rapid growth in exports could have been sustained without continuing liberalization in other markets is doubtful, at best.[2]

Seen in this light, a successful export oriented set of trade policies forces the adoption of other efficiency and growth-enhancing liberalization policies. Those policies permit further gains to be realized from the trade strategy, and simultaneously induce further growth and yet greater complexity of the economic structure. The costs of remaining interventions (credit rationing, regulations governing international financial transactions, etc.) then rise further as output increases. The "virtuous" circle then persists. The success of an outer-oriented trade strategy provides the momentum and impetus for further liberalization, which then permits further economic gains from the trade strategy.

Whether the "virtuous policy circle" accounts for the high overall economic growth resulting from the trade strategy adopted by the East Asian countries, or whether other factors are the key explanatory variables is worthy of further research. What seems to be certain is that the South Asian countries that adopted policies of import substitution and industrialization behind high walls of protection failed to adopt policies conducive to rapid economic growth.

---

[2] It is an interesting question as to whether costly interventions in agriculture in Korea and Taiwan will come to represent a major bottleneck to growth.

External Sector in Development: China and India, 1950–89

By T. N. SRINIVASAN*

China and India, the two continent-sized ancient societies, initiated their planning for national development at the same time. China's First Five-Year Plan covered 1952–57, while India's covered 1951–56. China is governed by a dictatorship and its economy (until the reforms of 1978) was a command economy. India is a representative democracy with an economy in which a large private sector and markets function (albeit tempered by government interventions). In spite of these differences, the development strategies in general, and foreign trade policies in particular, of the two economies are similar.

China and India were both influenced by the perceived success of the Soviet Union in rapidly industrializing a largely rural economy in a short period of four decades without significant external assistance. Both adopted a Soviet-style inward-oriented development strategy that emphasized heavy industry, producing intermediate and capital goods, though over time and with accumulating experience some changes in emphasis in the strategy have taken place.

The growth of per capita income is similar in the two economies. According to the World Bank, in 1952 U.S. dollars, China's per capita GNP in 1952 was $50, while India's in 1950 was $60. In 1987, at current U.S. dollars, the figures were $300 and $290, respectively. China's industrial value-added grew more rapidly at 12.3 percent per year than India's at about 5 percent per year. The emphasis on capital-intensive industrialization in both countries slowed the rate of growth of employment in industry so that the share of labour force in agriculture in 1980 was 74 percent in China and 70 percent in India.

In the postreform period of 1980–87, value-added by agriculture and total GDP in China grew at impressive rates of 7.4 and 10.4 percent per year, respectively. During the same period there were hesitant attempts at liberalization in India, though the growth of GDP at 4.6 percent per year and industrial value-added by 7.2 percent were less impressive.

Until the reforms, foreign trade activities were monopolized in China by nine national foreign trade corporations. Any profit or loss of these corporations were absorbed by the exchequer. Neither the producers nor the corporations had an incentive to be cost-conscious and efficient (Yun-Wing Sung, 1989). In India, since foreign trade was largely in the hands of the private sector, an elaborate system of exchange controls and allocation was instituted to ensure that the foreign exchange earned by exporters was used to import only those commodities that conformed to the priorities set in the five-year plans. The complexity of the system led to a wide range of implicit exchange rates across the spectrum of imports and exports and associated efficiency losses.

Because of their inward-oriented development strategies, both China and India failed to take advantage of opportunities offered by growing world trade. China's share of world exports fell from about 1.25 percent during 1952–55 to 0.75 percent in 1978 (Sung). India's share fell from about 1.50 during 1952–55 to 0.57 percent in 1978 (S. Bhalla, 1989). After the reforms, China increased its share substantially to 1.70 percent in 1989, while India merely stabilized the share around 0.52 percent during 1980–88.

---

*Samuel C. Park, Jr. Professor of Economics, Yale University, New Haven, CT 06520. I thank Yun-Wing Sung for allowing me to draw on his unpublished research on China. Most of the data cited in the paper have been drawn from the publications of the World Bank.

I. Foreign Trade Policy of China: 1950–89

Prior to 1978, foreign trade policy shifted course back and forth with the ideological wind. Foreign trade was not exploited to achieve efficient development based on China's dynamic comparative advantage. Three broad subperiods could be distinguished: the period before the onset of the Great Leap Forward in 1958, the period after the Great Leap Forward till the dust from the cultural revolution settled in 1970, and the period of recovery from the aftermaths of the cultural revolution till the overthrow of the Gang of Four.

Industrial output grew rapidly during 1952–57 resulting in massive imports of plant and equipment. Exports at constant world prices grew at the rate of 9.7 percent per year during 1950–60 and imports at 10.7 percent. Over 70 percent of the value of exports during 1952–59 consisted of food and crude materials including cotton.

Soviet aid ceased in 1960. The failures of the Great Leap Forward turned China from a modest net exporter of grain to a net importer. The share of food in exports dropped dramatically and that of manufactures rose to around 50 percent, where it remained on average for more than two decades from 1959. Imports of machinery had to be cut sharply. During the period 1963–65, the economy recovered from the ravages of the Great Leap and a famine that killed more than 20 million people. By 1966, the nominal value of trade regained its 1959 value. But Chairman Mao unleashed the cultural revolution and the Red Guards the same year.

The years 1966–69 were disastrous for the economy and society. The period 1970–79 saw the revival of foreign trade with the share of exports in GDP rising from 2.55 percent in 1970 to 5.46 percent in 1979. Foreign trade was no longer viewed as a necessary evil, but as an essential ingredient in modernizing the economy. Although the open-door policy and a modernization program initiated by Mao's successor Hua Guofeng were not well conceived and largely failed, they led to much more carefully thought-out reforms. These included the creation of new agencies to promote exports and attract foreign investment, and to introduce foreign technology. Also, foreign trade reforms were coordinated with domestic economic reform. However, progress followed "a two steps forward and one step backward" sequence. In each reform drive several areas were opened up and given autonomy in their foreign trade and investment decisions. This led to an investment and import boom resulting in inflation and balance of payments problems. These problems led to a pause soon to be followed by opening up more areas and so on. There were three such drives during 1979–88.

The open-door policy led to rapidly rising exports, particularly of manufactures. Total exports (at current prices) grew at an annual average rate of 12.8 percent during 1980–88, with the share of manufactured exports rising from 50 percent in 1980 to over 70 percent in 1988. Imports grew at an annual average rate of 13.5 percent. It is virtually impossible to evaluate the resource cost of this impressive performance. Substantial investment has gone into developing the infrastructure in the open areas, particularly in special economic zones near Hong Kong and Macau, consisting of highways, railway lines, housing, water supply, telephone systems, not to mention luxury hotels. A hard-headed cost-benefit analysis using appropriate prices and discount rates is needed to judge whether the return from this investment was as high as from other investment opportunities.

II. Foreign Trade Policy of India: 1950–89

India's development strategy has three broad objectives: economic growth, self-reliance, and social justice. It assigned a dominant role for the state in development with key industries such as railways, telecommunications, and electricity generation being the exclusive responsibility of the public sector. The planned massive step-up in investment during the second plan and its high import intensity created the expectation that foreign exchange would become scarce in the future, and precipitated a severe balance of payments crisis in 1957. Ever since, Indian planning has been driven by the perceived need to conserve foreign exchange expenditures. The broad features of the foreign exchange and investment licensing system instituted in the late 1950s has remained essentially unchanged, though its severity has waxed and waned.

The licensing mechanism was designed to approve (sometimes with modifications) or deny an application for license. Since the criteria for evaluation of each application were sufficiently broad, there was significant room for discretion on the part of the licensing authorities. This meant at best the licensing bureaucrats were under constant pressure from interested parties, and at worst led to their corruption and politicization of their decisions. The information system needed to keep track of the licenses granted and their implementation was not there. In the final analysis, the procedure for license granting followed rules of thumb rather than any economic rationale. It degenerated into an exercise in patronage (political and other) dispensation and a source of rents for personal or political use.

The following subperiods can be distinguished within 1957–89: the period starting from the foreign exchange crisis of 1957 and ending with the devaluation of the rupee in June 1966; the period 1966–68 covering an abortive attempt at economic liberalization; reversion to a restrictive regime during 1968–75; a decade (1975–85) of selective relaxation of controls in response to an exogenous improvement in foreign exchange availability, and finally, a period of more systematic and hopefully irreversible liberalization starting in 1985 (see Jagdish Bhagwati and myself, 1975, and my 1989 paper). The system of quantitative restrictions (QR) on foreign trade was also used to provide automatic and custom-made protection to any domestic activity that substituted imports. The realization that the QR regime was penalizing exports led to the introduction of a plethora of export subsidization schemes in 1962.

A severe drought in 1965 and the suspension of foreign aid following the Indo-Pakistan war of late 1965 led to a rise of import premia to unprecedented levels by early 1966. The resumption of large-scale significant aid as well as access to IMF credit was made conditional on the liberalization of the economy, the first step of which was the devaluation of the rupee. The government devalued the rupee in June 1966. Devaluation was also accompanied by an elimination of export subsidies and reduction of import duties.

The drought of 1965 was followed by another disastrous drought in 1966 and this resulted in price increases, adverse impact on traditional exports, and an industrial recession induced by shortage of agro-based raw materials. Devaluation was seen as capitulation to external pressure, not only by the opposition but also by some within the ruling Congress Party whose power was eroded in the 1967 general elections. Further, the promised external aid did not materialize. The hesitant steps towards liberalization introduced with devaluation were largely abandoned by 1968.

With the abandonment of liberalization, import premia were back at significant levels. Export subsidies were reinstated and augmented. Industrial licensing reverted to its severely restrictive mode. Import allocation criteria became increasingly complex.

The restrictive regime continued until the middle 1970s when foreign exchange availability improved dramatically, in part fueled by the increase in the number of and remittances from Indians working in West Asia. With the spread of the green revolution, food grain imports ceased. The net result of all this was a relaxation of the severity of the QR regime. Import allocation rules were made simpler, and allocation of foreign exchange for essential imports not competing with domestic production was increased.

The next stage in liberalization came in April 1985 when the government announced an import and export policy that was to cover a period of three years, rather than the six months or year that they did in the past. This innovation was meant to bring some stability to the policy. The latest (1988–91) policy statement candidly recognizes that changes have to be announced within the three-year period and states "it is proposed to issue a revised version of the import and export policy at the beginning of each year." So much for stability of policy!

The import control regime has confined imports to essential consumer goods, raw materials, and investment goods needed for domestic production and exports. The composition of India's exports has shifted away from primary products to manufactured goods, but only moderately, the share of the latter rising from about 50 percent in 1950–51 to as little over 74 percent in 1986–87. Yet, within the group of manufactured products, just four items (leather, gems, garments and textiles) account for 70 percent of the growth during 1980–88. A diversified and dynamic export sector is yet to emerge.

III. Lessons from China and India

The development strategy of India and China has restricted foreign trade beyond its relatively restricted role expected in large economies as compared to small open economies. The policy instruments used to implement this strategy created avoidable efficiency losses, including those arising from diversion of productive resources to rent seeking.

Both economies have succeeded in building a well-diversified industrial structure. Both produce a far greater variety of industrial goods including capital goods than most developing countries. They depend to a much lesser extent on imported equipment. Their scientific and technological capability is evident in their success in ventures ranging from putting satellites into orbit to designing and operating nuclear power plants. Both have so far avoided accumulating a heavy foreign debt, although India may be close to joining the rank of the heavily indebted. China has succeeded in eliminating abject poverty, improved the education, health and nutritional status of its population as evidenced by the high literacy rate (male and female) and high life expectancy at birth, though regional inequalities in income remain and may even be increasing after the reforms. Although India's achievements are less impressive, they are nonetheless significant.

These achievements have been obtained at a high cost. The diversified industrial structure is, in most if not all cases, not internationally competitive in cost and quality of output, certainly in India and perhaps in China. The Indian industrial and input licensing system encouraged the creation of small-scale, high-cost plants in many industries and sheltered them from internal and external competition. Most of the industries in the public sector have yielded negligible returns. It is doubtful whether the protected capital goods industries in the two countries have accumulated the dynamic learning experience to bring about significant technical improvements in equipments they produce. There is ample evidence showing that total factor productivity declined in India's manufacturing sector.

A comparison with South Korea will illustrate the extent of the failure of the two giants. In 1965, India (China) exported 8 (10) times as much of manufactured goods in value as Korea, and the value-added by India's manufacturing sector was more than 5 times that of Korea. By 1986, Korea was exporting more than 4.5 (1.5) times compared to India (China) of manufactures, and the size of its manufacturing sector was nearly two-thirds (one-fourth) of India's (China's). Korea's total external debt at the end of 1987 was $40.5 billion compared to India's $46.4 billion and China's $30.2 billion. Because of its much better GNP and export growth, Korea is not deemed a potential problem debtor while India is likely to be. The achievements of Sri Lanka, for example, with respect to life expectancy, literacy, etc. is superior to China's and far superior to India's.

The lesson is *not* that the Korean success is because the Korean government did not intervene in markets while those of China and India did. It is that the interventions in the Korean case did not distort incentives as much as those in India did. It is unlikely that India will achieve efficient and rapid growth (or that China will be able to sustain its recent rapid growth), unless there is vigorous competition in domestic markets. A dismantling of the Indian resource allocation system based on quantitative restrictions administered on a case-by-case basis, a thorough-going reform of the state enterprise and price system in China, and a greater reliance on world market signals through outward orientation in both countries are essential for invigorating dynamic competition. The achievements of Sri Lanka (at least until the recent ethnic conflicts) suggest that neither high per capita income nor sacrifice of freedom as in China is necessary to bring about improvements in basic quality of life of the entire population.

REFERENCES

Bhagwati, Jagdish and Srinivasan, T. N., *Foreign Trade Regimes and Economic Development: India*, New York: Columbia University Press, 1975.

Bhalla, S., "Indian Exports, Imports and Exchange Rates: A Comparative Quantitative Analysis," Policy Group, New Delhi, 1989.

Srinivasan, T. N., "India's Foreign Trade Policy and the Political Economy of Development Planning," Economic Growth Center, Yale University, 1989.

Sung, Yun-Wing, "An Appraisal of China's Foreign Trade Policy, 1950-1989," Economic Growth Centre, Yale University, 1989.

Development Lessons from Asia: The Role of Government in South Korea and Taiwan

By YUNG CHUL PARK*

South Korea and Taiwan share common structural characteristics and export-led industrialization. Several studies have documented an active role of government, beyond correcting market failures (H. Pack and W. E. Westphal, 1986; Paul Kuznets, 1988). But the role of government differs. The Taiwan government has been supportive rather than interventionist, whereas the Korean government has been collaborative and even coercive in relations with the private sector.

Government in both countries helped launch and sustain the export-led strategy, taking into consideration the initial conditions in industrial organization and structure, capital markets and trade. Although the strategy contributed to export development, once put in place, the respective governments have been unwilling and unable to adjust their role to changing circumstances. This rigidity accounts for the inefficiency of government in Korea in the 1970s and in Taiwan in the 1980s.

I. Contrasts

Since the import substitution in the 1950s, the Korean government has intervened in the allocation of resources creating an entrenched policy bureaucracy that controls the means for plan implementation. The institutional setup provides domestic market protection, implements industrial targeting, and issues permits of entry in many industries and foreign loan guarantees. "Korea, Inc." is a more apt description of government-business relations in Korea than Japan (Kuznets).

The Taiwan government uses a medium-term economic plan that sets macroeconomic targets and formulates sectoral investment plans. Unlike Korea, however, these plans are indicative. The planners have had little authority and limited access to policy instruments such as credit allocation. There is no institutional network linking the economic bureaucracy with the private sector. Interaction has been characterized by mutual adjustment rather than collaboration. The Taiwan government has in general confined its role to providing social and physical infrastructure and other public goods. It has balanced public needs with a desire to encourage private enterprise in its intervention.

II. Leading Role in Korea

In the 1960s, Korea had to cope with a formidable array of structural and institutional problems. The low rate of savings and chronic current account deficits required a continuous inflow of foreign lending unavailable without government guarantees. It was believed that the financial system could not direct resources to export-oriented industries and that there was a shortage of entrepreneurs and managers to undertake the development of export-oriented industries. The industrialist class that had been nurtured during the import substitution period had little knowledge of foreign markets and international marketing, and hence was unprepared to take high risks in selling abroad.

In many cases efficiency required adoption of increasing returns technologies. This conflicted with Korea's limited availability of resources, and forced the government to support a few, selected large producers in the targeted industries. In the absence of a market mechanism, they assumed the task of selecting the winners. The banking system was used to channel domestic and foreign savings to these large firms and as a de facto partner, and the government was drawn to participate in their business decisions. During the 1960s, these large firms became successful exporters, and with the growth of the economy also developed into industrial groups dominating the manufacturing sector.

---

*Professor of economics, Korean University, Seoul, Korea.

Korea's export-oriented industries could have sustained rapid growth without government intervention, though not without government support in the 1970s. However, instead of deregulating the controlled sectors, Korean policymakers tightened their grip over manufacturing industries and financial intermediaries. The confidence they gained from the success in the 1960s encouraged a big, unsuccessful import substitution for capital and technology-intensive products in the 1970s.

Industrial concentration and government involvement created its own problems. By 1977, the 30 largest groups accounted for 25 percent of employment and 34 percent of shipment of manufacturing (Kyu-uck Lee, 1986, p. 239). Concentration of economic power in the hands of a few conglomerates was necessary for efficiency, but highly undesirable for distributive equity. This dilemma created pressure for (and justified the government control of) the industrial groups. The pricing and supply behavior of these groups had to be regulated because their market position was protected by closed trade and a closed financial regime.

Government support for industrial groups created moral hazard problems. By the early 1970s, they were highly leveraged with loan guarantees through the banks owned by the government. The government had literally become a partner responsible for their failure as well as success. This bailout certainty induced excessive risk taking.

The coordination and moral hazard problems were most serious during the 1970s when the government attempted to develop simultaneously a number of heavy and chemical industries. The development was deemed necessary to take advantage of complementarity on the demand side, and forward and backward linkages among these industries. The industrial groups lost no time in making massive investments, not to be locked out of potentially lucrative markets. The result of this competitive bidding was an excessive investment and duplication.

Except for petro chemicals and nonferrous metals, most industries were developed as future export industries. In order to export with the benefits of scale economies and minimum efficient size, the groups were encouraged to build large plants from the beginning. But because of marketing, quality, and technology problems, export earnings were low and losses heavy. The complementarity as well as forward and backward linkages made the situation much worse. This setback motivated economic liberalization at the end of the 1970s. Yet, after almost a decade of trade and financial liberalization in the 1980s, Korea's policy regime is still based on development mercantilism. The government has not been able to extricate itself from its relations with the industrial groups.

III. Supportive Role in Taiwan

The indirect role of government in Taiwan was predicated on a high rate of savings, chronic trade surpluses, conservative stance of fiscal and monetary policy, an egalitarian development philosophy, and an industrial organization characterized by a large number of small and medium-sized firms in manufacturing. As a result, Taiwanese planners have been largely deprived of two instruments of industrial policy—the control over credit and budgetary allocation for development purposes.

The political leadership was also determined to prevent the concentration of resources in private hands, as it believed that a small group of powerful businessmen who controlled a large share of private resources and the financial system had caused its downfall on the mainland. This determination was manifest in the proliferation of public enterprises that accounted for an average 50 percent of manufacturing value-added during the 1950s. Even in the export promotion strategy led primarily by the private sector, this egalitarian tradition remained a balancing force.

In Taiwan, a large number of small and medium-sized firms produce and export a large share of Taiwan's manufactured products. In 1986, for example, more than 98 percent of Taiwan's firms had fewer than 300 employees, and 48 percent of these employed 5 or less workers. Small and medium-sized firms produced about 30 percent of manufacturing value-added, and exported most of this to foreign markets (Tyler Biggs, 1988). Taiwan embarked on export promotion of labor intensive manufactures just as Korea had, but chose different production technologies. Taiwan's planners did not select increasing return technologies as the Korean planners did, because they could draw on a large pool of experienced entrepreneurs.

The dominance of small and medium-sized firms continues in Taiwan's major export industries. For example, in the footwear industry, firms with 300 or less workers accounted for almost 60 percent of value-added in Taiwan, whereas the same share was about 7 percent in Korea in 1976 (Brian Levy, 1987). With this structure, direct intervention Korean-style was impossible. A more efficient way of supporting these small exporters was to provide uniform incentives on the basis of export performance.

The strategy of relying on small and medium-sized firms obviously raises problems related to the loss of scale economies, marketing and collection, and dissemination of information on foreign consumers' preferences and technological developments. But in Taiwan they were overcome by the presence of a large number of traders coming from the mainland (see Levy). These traders subdivided large orders among small producers and still do. They gave Taiwanese manufacturers access to orders of small volume that Korean manufacturers could not accept. They have also successfully sought out specialized products for which demand is relatively small, and many technology-intensive activities to which small firms are well adapted (Levy and Wen-Jeng Kuo, 1987). Perhaps this lack of success was fortunate in that it avoided the large-scale inefficiency associated with the Korean development effort in heavy industry.

Unlike Korea, Taiwan's efforts at import substitution of capital and intermediate goods in the 1970s were frustrated. Entrepreneurs were unable and unwilling to make large capital investments with a long gestation period. Neither did the government have an industrial policy regime capable of planning and implementing an import substitution program as ambitious as Korea's scheme. In order to bypass these difficulties, the government entrusted public enterprises with developing the automotive industry in the 1970s, but the attempt did not succeed (Yun-Han Chu, 1989).

Since the early 1980s, the Taiwan government has widened the scope and accelerated the pace of trade and financial liberalization. However, economic liberalization has not mitigated the imbalance between the traded and nontraded goods sectors. The traded goods sector is still too large, the nontraded goods sector too small. An expansion of domestic demand, in particular in the public sector, is now needed. Yet, the industrial policy regime that prevented the promotion of a large-scale import substitution plan in the 1970s has made it equally difficult to channel more resources into the nontraded goods sector. The housing and social and physical infrastructure sectors remain grossly underdeveloped. The supportive role of government has become a cause of misallocation of resources.

IV. Any Lessons?

One conclusion is that there is no ideal model for the role of government. The success or failure of export-led development strategy does not seem to correlate with policy activism. In both Korea and Taiwan, initial conditions, structural, and institutional characteristics have been more important than the strategy itself, or development ideologies in shaping the role of government. Institutional and structural characteristics explain why Korean planners have always sought to adopt increasing return technologies whereas Taiwanese policymakers have not. Technology differences in turn brought about differences in the role of government between the two countries.

Another lesson is that once government assumes a certain role in the development process, it tends to persist with the same role, often failing to adjust to changes in economic environment. This inflexibility could be a cause of government failure as it was in the case of Korea in the 1970s and in the 1980s in Taiwan.

Finally, have governments contributed to rapid growth and industrialization in the two countries through intervention? It is difficult to believe that the private sector alone could have launched and sustained the export drive without the direct government intervention in Korea in the 1960s. Nor would it have been more efficient if the Taiwan government had assumed a more interventionist role during the same period. But since the mid-1970s, efficiency of government intervention in the two economies has been debatable at best. Judging from the record of rapid growth with stability, however, any losses related to government intervention must have had their offsets.

REFERENCES

Biggs, Tyler S., "Financing the Emergence of Small and Medium Enterprise in Taiwan: Financial Mobilization and the Flow of Domestic Credit to the Private Sector," Employment and Enterprise Development Division Discussion Paper No. 15, USAID, Washington, August 1988.

Chu, Yun-Han, "The State and the Development of Automobile Industry in South Korea and Taiwan," presented to Conference on State Policy and Economic Development, Taiwan, December 1989.

Kuznets, Paul W., "An East Asian Model of Economic Development: Japan, Taiwan, and South Korea," *Economic Development and Cultural Change*, April 1988, Suppl. *36*, S11–S44.

Lee, Kyu-uck, "The Concentration of Economic Power in Korea: Causes, Consequences and Policy," in his *Industrial Development Policies and Issues*, Seoul: Korea Development Institute, 1986.

Levy, Brian, "Export Intermediation and Structure of Industry in Korea and Taiwan," Korea Development Institute Working Paper No. 8717, December 1987.

______ **and Kuo, Wen-Jeng,** "Investment Requirements and the Participation of Korean and Taiwanese Firms in Technology Intensive Industries," Korea Development Working Paper No. 8718, December 1987.

Pack, H. and Westphal, W. E., "Industrial Strategy and Technological Change: Theory versus Reality," *Journal of Economic Development*, June 1986, *11*, 87–128.

INCENTIVE EFFECTS OF MEDICAL MALPRACTICE[†]

The Effects of Malpractice Litigation on Physicians' Fees and Incomes

By PATRICIA M. DANZON, MARK V. PAULY, AND RAYNARD S. KINGTON*

For many years physicians and other medical providers in the United States have been subject to a negligence rule of liability. Under a negligence rule, patients who suffer an adverse outcome are entitled to compensation if they can show that they incurred an injury that was caused by the physician's failure to take due care, defined as departure from the customary standard of care of physicians in good standing in the profession. The traditional tort standard of damages is full compensation for monetary and nonmonetary loss.

There have been no major doctrinal shifts in the law expanding liability for medical malpractice over the last decade—indeed, many states have enacted tort reforms designed to reduce the number and size of claims. Nevertheless, from 1975 to 1984, claims per physician rose at an average rate of 10 percent a year; between 1982 and 1986 claim frequency per 100 physicians rose from 13.5 to 17.2 a year. (See Danzon, 1988, and sources cited therein.) Claim severity (average amount per paid claim) increased at roughly twice the rate of the CPI from 1975 to 1984. In 1984, the median and mean payment were $18,000 and $80,741, respectively. There remain large differences among states and among specialties in claim frequency and severity.

Costs of malpractice insurance have also increased dramatically, but more erratically than claims costs. Following the malpractice "crisis" of the mid-1970s, when rates increased by over 300 percent in a single year in some states, rates remained stable or actually fell in the late 1970s, but resumed an upward trend in the 1980s. Between 1977 and 1984, rates for basic limits of coverage increased by 109 percent for the lowest-risk specialists, such as GPs doing no surgery, 180 percent for obstetricians, and 189 percent for the highest-risk surgical specialists.

Although malpractice insurance is still less than 2 percent of total health care expenditures, many observers argue that medical malpractice is a major factor contributing to rising health care costs. One implication of such arguments is that virtually all the cost is borne by patients. This seems at odds with arguments made by others, that a significant fraction of physicians are giving up practice, or at least high-risk procedures, because of liability.

Full and automatic pass-through of costs is also potentially at odds with the deterrent purposes of the negligence rule of liability, which is to correct the potential underprovision of care in a market with asymmetric information and systematic underestimate of risk by consumers. With a shift from no liability to a perfect negligence system, we might observe an increase in preventive measures by physicians, an increase in prices, and an increase in gross revenues. This would occur *if* patients perceived the improvement in quality and valued it at cost, and the

---

[†] *Discussants*: W. Kip Vicusi, Duke University; Charles Phelps, University of Rochester.

*Department of Health Care Systems, The Wharton School, and Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA 19104. This research was supported by a grant from the Robert Wood Johnson Foundation Program to the Leonard Davis Institute of Health Economics at the University of Pennsylvania. Danzon was also supported by the Lilly Endowment, through the Center for the Study of the Economy and the State at the University of Chicago. This support is gratefully acknowledged, as is very able research assistance by Paul Kumagai.

system had been nonoptimal without liability—but nonoptimality is inconsistent with fully informed patients.[1] With imperfect information on the part of courts, liability insurers, and consumers, the incidence and incentive effects of the tort system are uncertain a priori and become an empirical question.

In this paper we discuss some of the theoretical and empirical issues in estimating the effects of malpractice litigation on physicians' fees and incomes and summarize some new empirical estimates.

I. Effects of Malpractice on Physician Pricing in the Short Run

Assume that cross-sectional differences and the intertemporal increase in claim frequency and severity reflect changes in the legal system that increase the probability and expected size of a malpractice recovery for the plaintiff, conditional on the occurrence of an injury.[2] Let us characterize the "generosity" of the legal regime by an index J.

Malpractice litigation creates two types of cost for physicians: the cost of claims and legal defense that can be covered by insurance; and the cost of own time in claim defense, loss of reputation, anxiety, etc. that are uninsurable. Under current methods of rating malpractice insurance, the cost of insurance depends on the physician's specialty, limits of coverage, whether or not certain high-risk procedures are performed at all (for example, any surgery by GPs, obstetrics by OB/GYNs) and the decision to practice full or part-time. Given these basic practice choices, the premium is a fixed cost, independent of number of patient contacts.[3] The uninsured costs are a marginal per patient cost equal to the expected cost if sued. This marginal cost increases with claim frequency and severity. On the other hand, if physicians maximize utility rather than profit, an increase in premiums that reduces net income should lower the physician's implicit opportunity cost of time, assuming leisure is a normal good. In that case, an increase in J should lower marginal cost. An increase in J may also affect patients' demand for care, if patients perceive and value any risk-reducing measures taken by physicians, or if they perceive and value any increase in the probability or amount of compensation if injured.

How does an increase in J affect physicians' prices and net incomes? Consider the simple case where firm-level demand is not affected by changes in potential malpractice awards and the physician-firm is a profit maximizer operating in a monopolistically competitive market (Pauly and Mark Satterthwaite, 1981). Further assume that long-run equilibrium requires the equalization of (real) net incomes across market areas.

We consider two basic models of the process of adjustment to market equilibrium following an increase in J. At one extreme, physicians simply purchase higher limits of coverage but do not otherwise change their practice patterns. There is full incidence of the increased insurance and uninsured costs on physicians and hence a sharp reduction in net incomes and even larger reduction in utility. This would be a short-run equilibrium if physicians are profit maximizers and the changes in premiums are viewed as a fixed cost.

---

[1] If patients perceive the average quality of care but not the care of individual physicians, then full passthrough of the costs of additional care under a perfect negligence rule may be possible, even though market forces alone would create suboptimal incentives for care.

[2] For an analysis of the effects of legal, medical, and demographic factors on the frequency and severity of malpractice claims by state, see Danzon (1984; 1986). California data for 1974 indicate that, at most, 1 in 10 injuries due to negligence led to a claim and 1 in 25 received compensation. Thus some increase in claim frequency is possible without assuming a change in the rate of negligence or a proplaintiff shift in the law.

[3] Although explicit experience rating is not widespread, implicit experience rating occurs because physicians with very poor claims records may face restrictions on their practice or be denied coverage by more selective insurers who charge lower premiums. Periods of sharply rising premiums marketwide are usually accompanied by a tightening of underwriting standards.

In long-run equilibrium, stable differences in J must be reflected either in price levels or in differences in "quality" that are cost saving to physicians, assuming equalization of (real) net income per physician across market areas. Physician migration is expected to be the primary equilibrating device in this first model if premiums are treated as a fixed cost by individual profit-maximizing physicians and increases in malpractice litigation do not raise marginal costs due to (implicit) expected uninsured costs.

In the second model, an increase in J does significantly raise physicians' perceived marginal costs and trigger an increase in prices in the short run that could exceed the increase in marginal cost.[4] Assuming that the market for physicians' services is monopolistically competitive, the profit-maximizing change in price is

$$(1)\quad dP/dJ = [dMC/dJ][1+(1/ED)]^{-1}$$

where MC is marginal cost and ED is the firm-level price elasticity of demand, assumed constant. Price may increase by more or less than MC depending on ED. For example, using Thomas McCarthy's (1985) rough estimate of -3 for demand elasticity, price will rise by 3/2 times any increase in marginal cost. However, if demand becomes more elastic as price rises (for example, if demand is linear), then price may rise by an amount less than or equal to marginal cost.

But, if the increase in J affects all physicians in an area equally, then the final effect on prices and incomes depends on market demand, not the individual firm demand elasticity. If the market demand elasticity is low, the final outcome can be an increase in price with minimal reduction in quantity, almost no change in gross or net income, and hence almost no migration. This combination of relatively elastic demand at the firm level but inelastic demand at the market level is a reasonable conjecture for physician services markets.

There are alternative assumptions that would yield the same short-run effect on price. The physician services market might be characterized as oligopolistic, and the increase in fixed costs might be used as a signal to raise prices. Alternatively, if insurers rapidly and automatically incorporate premium cost increases into their reimbursement levels, institutional mechanisms could facilitate a rapid adjustment in fee levels such that physician incomes are protected from even short-run incidence and there is no strong stimulus to adjustment of physician stocks.

II. Data

The data used in this study are from the 1976, 1978, and 1983 surveys of Physicians' Practice Costs and Incomes (PPCI), conducted for the Health Care Financing Administration. Each of these surveys occurred at a different point in the malpractice insurance "cycle." The 1976 survey follows two years of rapidly rising premium rates; in the two years prior to the 1978 survey, rates were flat or falling. Beginning in 1980, rates then began to rise again.[5] These data permit both cross-sectional and time-series (first difference) analysis of the relationship between physicians' fees, reimbursement levels, and incomes, on the one hand, and measures of the malpractice climate on the other.

We use three measures of the "malpractice climate": state-level claim frequency per physician, state-level claim severity (average payment per paid claim), and the rate charged by the leading malpractice insurer in the state for basic limits coverage. This rate is not a fully accurate measure of the relevant price of insurance because the majority of physicians buy excess limits of coverage, for which rates are nonlinear, but fairly uniform across states. Claim frequency is a proxy for the uninsured time costs of being sued. Potential claim severity would be a measure of exposure in excess of basic limits, but is imperfectly measured by *observed* severity because the latter tends to be inversely correlated with claim frequency. Controlling for claim frequency and claim severity, the premium rate is a proxy for the loading charge for claim defense, administration, and profit, and any discrepancy between prior claim frequency and severity and expected future claim costs. Since all three measures are potential proxies for different dimensions of the malpractice climate, all three are included in the empirical analysis reported here. An alternative specification uses the individual physician's actual expenditure on malpractice insurance, treated as endogenous. The analysis controls for other relevant market area characteristics that are expected to affect physicians' fees and incomes.

---

[4] Another alternative not addressed here is that physician outflow may reduce physician inputs and quality per unit of output, or some combination of price increases and quality decreases.

[5] The later 1986 PPCI survey follows the very sharp premium increases of the mid-1980s, but unfortunately does not contain data on prices.

III. Empirical Results:

A. *Cross Section*

In all three years, the cross-section evidence shows a strong positive relationship between the various measures of the malpractice climate and physicians' fees (Table 1).[6] The elasticity with respect to the physician's (endogenous) expenditure on malpractice insurance is roughly .16 for office visit fees, and between .09 and .17 for hospital visit fees. The pattern across years does not support the hypothesis of delayed response of fees. This hypothesis would predict a larger positive relationship between fees and rates in 1978 (when physicians had several years to adjust prices to the sharp rate increases in 1974–76) than in 1976 or 1983 which immediately followed sharp rate increases. Controlling for the rate for basic limits coverage, high claim severity is associated with higher fees, whereas the coefficient on claim frequency is generally negative or insignificant for office visit fees, and positive for hospital visit fees.

TABLE 1—CROSS-SECTION RELATIONSHIP BETWEEN MALPRACTICE MEASURES AND FEES

Variable	1976	1978	1983
A. Dependent variable: Usual charge for office visit (Ln)			
RATE (Ln)	0.105	0.103	0.113
	(10.22)	(10.49)	(6.74)
CLAIM SEVERITY (Ln)	0.028	0.040	0.022
	(1.88)	(3.47)	(2.14)
CLAIM FREQUENCY	−0.006	0.003	−0.113
	(−1.70)	(0.80)	(−3.16)
PREMIUM[a] (Ln)	0.155	0.165	0.176
	(8.82)	(9.52)	(6.52)
B. Dependent variable: Usual charge for hospital follow-up visit (Ln)			
RATE (Ln)	0.039	0.047	0.110
	(2.98)	(3.71)	(4.46)
CLAIM SEVERITY (Ln)	0.087	0.060	0.015
	(4.23)	(3.91)	(0.99)
CLAIM FREQUENCY	0.014	0.01	0.016
	(2.91)	(2.12)	(0.31)
PREMIUM[a] (Ln)	0.140	0.085	0.173
	(5.98)	(3.96)	(4.59)

Note: Regression coefficients and *t*-statistics are shown in parentheses.

[a] Two-stage least squares.

The magnitude of the price elasticities suggests more than a full pass-through of the costs of insurance. There are several possible explanations for this. First, as discussed earlier, an increase in marginal cost can result in a more-than-proportional increase in price if market demand is inelastic and physicians act as profit-maximizing monopolistic competitors, or if third-party payers rapidly adjust reimbursement rates, as they appear to have done. Second, if physicians are utility maximizers and derive significant negative utility from the threat of being sued, they would require positive compensating income differentials to remain in practice. A third possibility is an increase in "quality," such as time per visit. Preliminary analysis does show some evidence of increased time per visit.

---

[6] Full equations are reported in our 1990 paper.

The elasticities of usual reimbursement paid by "the highest paying" Blue Shield and commercial insurer in the area and by Medicare are broadly similar to the elasticities for physicians' usual charges. However, the ability to pass-through malpractice costs to Medicaid reimbursement is more limited and appears to diminish over time, with a negative relationship by 1983. This is consistent with the general tightening of Medicaid limits on physician reimbursement in the 1980s.

When the sample is pooled across specialties, the relationship between malpractice variables and net incomes is consistently positive, although weaker in 1983 than in the earlier years. This effect largely disappears when the data are disaggregated by specialty, suggesting that pooling across specialties induces a spurious positive correlation between net income and the specialty-specific malpractice rate, even with specialty dummies included in the estimating equation.

These cross-sectional results are consistent with a model of the adjustment process in which there are significant positive and rapid adjustments in prices and reimbursement levels, and no significant negative effects on net income.

B. *Time-Series*

To distinguish between the models of adjustment via changes in physician stocks, on the one hand, and rapid adjustment of prices on the other hand, we performed two types of tests. First, we reestimated the cross-section equations treating county-level physician stocks as endogenous. The results were essentially unchanged. We also tested for the effect of *changes* in the malpractice variables on *changes* in physician stocks, and found no significant effects. Doctors did not leave (or choose to avoid) states whose malpractice climates worsened, relative to other states.

Second, we examined the relationship between changes in our measures of the malpractice threat and the contemporaneous change in net incomes, prices, and hours over the seven-year period, 1976–1983. Since the individual physician data are not in the form of a panel, we cannot examine first differences at the level of the individual practice. Instead, we calculate sample average values for observations grouped into the 72 malpractice insurance rating territories, and take first differences in those averages. Because the sample number of physicians per territory varies, results are reported using means weighted by the number of physicians in the territory. To control for changes in other area characteristics that might affect fee levels, we include the change in the predicted value of Medicare prevailing charges for four physician services: GP office visits and hospital visits, hysterectomies, and cholecsystectomies. Because of the exploratory nature of this analysis, we discuss the results of stepwise analysis, with a selection significance level of 0.15.

The first difference in malpractice premium rates is positively associated with the change in average fees for office visits and for follow-up hospital visits, with significance at better than the 0.1 level. Other significant covariates of fee changes have the expected signs. None of the malpractice variables was significant in explaining the change in net income or changes in insurers' reimbursement rates. For the 1976–78 period, during which malpractice premiums generally fell, the positive association between the malpractice variables and fees did not hold, suggesting that even if fees adjust rapidly upward, they are sticky downward. There was also no evidence of income decline associated with the malpractice variables.

In conclusion, the evidence from the 1976–83 period is consistent with the increase in malpractice costs being passed on fairly rapidly in higher fees, with little negative effect on physician net incomes on average or adjustment in physician stocks across market areas. We emphasize that these conclusions are based on data that predate the sharp increase in malpractice premiums in the mid-1980s. The pass-through of these later cost increases may have been much less rapid because of the more competitive market for physician services and more aggressive cost-control practices of third-party payers. Second, finding no evidence of effects on mean net incomes does not preclude significant effects for individual physicians and adjustment of physician stocks or service flows within insurance rating territories.

REFERENCES

Danzon, Patricia M., "Medical Malpractice Liability," in Robert E. Litan and Clifford Winston, eds., *Liability: Perspectives and Policy*, Washington: Brookings, 1988.

______, "The Frequency and Severity of Medical Malpractice Claims," *Journal of Law and Economics*, April 1984, *27*, 115–48.

______, "The Frequency and Severity of Medical Malpractice Claims: New Evidence," *Law and Contemporary Problems*, Spring 1986, *49*, 57–84.

Danzon, Patricia M., Pauly, Mark V. and Kington, Raynard S., "Incidence and Incentive Effects of Medical Malpractice," mimeo., University of Pennsylvania, 1990.

McCarthy, Thomas, "The Competitive Nature of the Primary-Care Physician Services Market," *Journal of Health Economics* June 1985, *4*, 93–118.

Pauly, Mark V. and Satterthwaite, Mark A., "The Pricing of Primary Care Physicians' Services: A Test of the Role of Consumer Information," *Bell Journal of Economics*, Autumn 1981, *12*, 408–506.

Experience Rating: Does It Make Sense for Medical Malpractice Insurance?

By FRANK A. SLOAN*

The tort system of which medical malpractice is part has two functions—to compensate victims of injuries due to negligence of another party, and to deter such negligence. Although the system seems to perform its compensation role poorly, there is little empirical evidence on the system's deterrent effects. Critics assert that the system overdeters, and the cost of such allocative inefficiency is substantial. Whatever the deterrent effect, there is plausibly less deterrence with complete third-party insurance with little or no experience rating of premiums. Although experience rating is frequently used in many lines of insurance, there has been considerable resistance to experience rating in the medical malpractice line.

Section I describes evidence on the prevalence of experience rating of medical malpractice insurance premiums and reasons given against such rating. Section II summarizes pertinent statistical evidence from past studies of experience rating in medical malpractice insurance and in other lines. Section III discusses options for experience rating at both individual physician and group levels. Finally, Section IV presents a summary and some conclusions. "Experience rating" refers to a variety of plans that base future premiums on past claims or loss experience of insureds. This study deals only tangentially with risk classification approaches based on other criteria.

I. Extent of Experience Rating in Medical Malpractice

During 1987–1988, Randall Bovbjerg, Penny Githens, and I (1990) conducted a survey of 14 medical malpractice insurers, 3 multiple-line, commercial stock, and 11 single-line, physician-sponsored mutuals or reciprocals specializing in medical malpractice insurance. The survey dealt with several topics of which experience rating was only one.

Only 1 of the 14 companies had never implemented an experience-rating program of any type. However, most had either completely abandoned experience rating or operated a limited program. Virtually all the plans were company initiated, rather than mandated by state statute or regulation.

The programs were limited along these dimensions. First, very few physician insureds were typically affected. Often, fewer than 1 percent of physician insureds paid more than standard rates because of adverse prior claims experience. The claims data base used for experience rating purposes was limited by one or a combination of several devices, such as by excluding claims with indemnity values below a certain dollar threshold and only counting claims that occurred recently. Often only recent claims were counted against the physician, since insurers believed that old claims have little predictive value. Insurers tended to use open claims for experience-rating purposes for the same reason. Medical malpractice claims often take many years to close. With open claims, however, physician liability has not yet been established and estimates of loss must be forecasted.

Second, none of the surcharges exceeded 200 percent of standard rates, and they were often much lower. Third, rather than apply a strict formula based on prior claims experience, the physician's record was often reviewed by a physician panel before a surcharge was applied with the option for the physician to appeal.

---

*Director, Health Policy Center, Vanderbilt University, Nashville, TN 37235. This research was supported in part by Grant No. 12412, "Scheduling Damages and Experience Rating: The Next Medical Malpractice Insurance Reform," from The Robert Wood Johnson Foundation. I have benefited from numerous conversations about experience rating with Randall R. Bovbjerg, Penny Githens, and Mahmud Hassan. In addition, I thank Gregory Alff, Kevin Thompson, and Jean Lemaire for advice on specific aspects of the study.

For purposes of computing experience-rated premiums, insurers differed in whether they counted claims filed or dollar amounts of loss. Claims frequency is used in experience-rating plans in automobile liability because claim size is thought to be beyond the policyholder's control. A careless driver presumably does not know in advance whether his victim will be a high- or low-wage earner. None of the respondents had "hard" empirical evidence that experience-rated premiums help to control frequency or severity of claims.

Several respondents expressed reservations about experience rating. Medical organizations and their physician constituents frequently opposed it. Surcharged physicians often left the company because they were able to secure lower premiums from competing insurers. This occurred because each insurer only used its own experience in premium setting. Experience rating was said not to work in competitive insurance markets because a surcharged physician could always find a lower-priced alternative.[1] Some respondents stated that experience rating is inappropriate in a line with a low-claims frequency such as medical malpractice insurance.

II. Statistical Evidence

Medical Malpractice Insurance. Losses in medical malpractice are highly concentrated. My coauthors and I (1989) found that 87 percent of losses of anesthesiologist and obstetrician-gynecologist policyholders in Florida were incurred by 6 percent of such physicians. Losses for internists and other surgical specialties were similarly concentrated. Also, controlling for specialty, past losses were highly predictive of future losses. Physician specialty is used by all insurers for risk classification purposes.

More formal treatments have analyzed claims frequency as a Poisson process (John Rolph, 1981; Blaine Nye and Alfred Hofflander, 1988; Randall Ellis et al., 1990; and myself and Hassan, 1989). More specifically, the claims frequency distribution in these studies was assumed to be Poisson distributed with mean λ. To recognize heterogeneity of claims experience within risk classes, λ was assumed to have a *gamma* distribution. Using this compound Poisson model and the method of moments to derive estimates of the underlying parameters, the studies obtained widely varying expected claims rates within risk classes (specialty).

For example, in Ellis et al., expected paid claims frequency typically rose steeply with number of paid claims in the five previous years. Using a method for accounting for severity (indemnity per paid claim) developed by Jean Lemaire (1985), Hassan and I found that, if anything, incorporating severity accentuated the effect of past experience on future premiums. Surcharges were as high as 400–500 percent. In sum, more accurate premiums would display much more dispersion than premiums do now.

Experience Rating in Other Lines. Premiums in several other lines of insurance are based on prior experience, and there is evidence from these fields that experience rating reduces claims frequency and injury rates.

Workers' Compensation. Experience rating is universal in workers' compensation. The experience rating unit is the individual firm. A firm-specific experience modification factor is multiplied by the rate for the industry class. The firm's own experience is given a higher weight when the experience is thought to reflect expected loss rather than random error (i.e., experience is more "credible"). The most recent evidence suggests that firms in states with high workers' compensation benefits, which are in turn exposed to high premiums, take more precautions to prevent work-related accidents (Michael Moore and W. Kip Viscusi, 1989).

Unemployment Insurance. Payroll taxes are used to finance unemployment insurance benefit payments. All states must have a UI program, and all states use experience rating. States have a choice of four methods for taxing firms (Robert Topel, 1984). Although firms' incentives differ depending on the precise method used, as a generalization, firms with consistently high separation rates receive subsidies from the pool which raises unemployment (again, see Topel).

---

[1] One would think that insurers would want to lose a policyholder who pays a premium less than expected loss.

Automobile Insurance. Automobile liability insurance is a high-frequency, low-severity line with a comparatively short claims tail. Experience rating is universal in North America, Western Europe, and Japan, and accident frequency appears to be lower in countries that use experience rating more extensively (Lemaire, 1988). Inferences about the effect of experience rating can also be made from empirical evidence on the impact of no-fault automobile insurance laws on motor vehicle injury rates. Elizabeth Landes (1982) found that injury rates involving automobiles increased appreciably in those states adopting no-fault between 1971 and 1975.

III. Alternatives for Experience Rating of Medical Insurance Premiums

A. *Choices in the Design of an Experience Rating Plan*

Mandatory vs. Voluntary Program. For government, there is the decision whether or not to require experience rating, and, if so, which uniform procedures to adopt. For various reasons, including medical society sponsorship of insurers, there may not be enough experience rating or careful underwriting in the absence of state action, thereby providing an insufficient incentive for accident prevention. Mispricing and careless underwriting may cause insolvency. Further, state government may collect claims data from the various insurers in the state. At the other extreme, too much experience rating may lead to excessive risk classification error (Thomas Palfrey and Chester Spatt, 1985). Although there is a state interest in quality assurance, solvency, and equity, public policymakers may not be more knowledgeable than private decision makers.

Experience Rating Unit. If individual medical malpractice premiums are based entirely on expected losses, the resulting surcharges and discounts tend to be substantial, which places the insureds at high risk of classification error. Basing premiums on group experience reduces this risk, but also blunts the potential deterrent effect of experience rating unless the group is given an incentive to police its members.

Time Period for Experience Rating. A decision must be made about the number of years' experience on which the premium is to be based, and the extent to which the experience with other companies and related information, such as disciplinary actions against the physician, is to be used. Also, insurers must decide whether information on open claims is to be used, and if so, how. Open claims have the advantage of more accurately reflecting current physician quality and practice patterns. However, their use requires that premiums be based on projected losses that may be considered to be "soft." A problem with the use of closed claims is that physician cooperation is needed for settlements to occur. To the extent that losses affect premiums, physicians may be more resistant to settling.

Exclusions. One reason for a deductible is to encourage settlement of small claims. Another is to encourage physicians to report incidents likely to result in claims, even if they seem minor. Small claims, if not attended to early, often grow into larger ones. On the other hand, excluding such claims from experience rating may raise frequency of minor mishaps.

Role of Peer Review. There are two issues. Should peer review be conducted and, if so, how should it be done? Claims are often settled even when liability is questionable. In one sense, one should base premiums on expected loss for whatever reason the loss occurs. Physicians with poor interpersonal skills, risky procedure mixes, and/or litigious patients should probably pay higher premiums. However, the case for some type of screening *is* strong when experience rating is based on open claims. Further, although it may be appropriate for physicians to charge suit-prone patients higher fees, this may not work in practice. Some health insurers may have the power to impose fixed fees on physicians. Also, it is often difficult to identify suit-prone patients in advance. Settlement should be easier if the insured physician knows that the case will be scrutinized by peer reviewers.

Medical society-run peer review may be anticompetitive. It also may be in physicians' collective interest and individual interests to undersupply effective peer review. Yet physicians may fear that peer review sponsored by insurers may be oriented toward dumping physicians with adverse claims experience, regardless of cause.

Relationship of Experience Rating to Other Policies. Insurers must decide on the appropriate mix of underwriting, experience rating, and other loss control activities. The actuarially appropriate schedule of surcharges-discounts based on the physician's own experience is substantially affected by the types of nonexperience-based risk classification system that is used (see Hassan and I). Premiums need not rise as much as a function of previous experience if other risk classification criteria are also used (for example, procedure mix, work hours).

B. *Group-Based Experience Rating*

Hospital Medical Staff. Rather than deal with many of the complex issues already identified, it may be preferable to base a premium on a group's experience and leave individual rating issues for the groups to solve. Virtually every hospital's medical staff engages in some peer review to determine medical staff membership and to provide some oversight of quality of care performed at the hospital. Such peer review, when conducted in the hospital's interest, is potentially procompetitive (James Blumstein and myself, 1988). Experience rating at the level of the hospital's medical staff would give peer reviewers an explicit financial incentive to conduct meaningful review. If physicians at a hospital were adverse to tough peer review, they would pay higher premiums. My paper with Hassan's empirical analysis of loss volatility of medical staffs of various sizes suggests that, given a sufficiently long record of past claims (5 or 6 years), it would be practical to rate staffs at even some small hospitals (particularly at those with favorable claims experience). Medical staffs that self-insure could reduce the risk of plan bankruptcy by reinsuring.

Group size is not the only issue, however. Some might object that staffs with sophisticated specialty mixes would pay higher premiums, but this is true to some extent now. Relativities in premiums across specialty-risk classes now in use could be employed to allocate premiums to individual members.

Although staffs would have a much greater incentive to exclude poor quality physicians than now, staffs might exclude physicians who treat patients prone to have adverse outcomes, or to charge such physicians fair premiums. Such physicians' premiums are cross subsidized now, but the cross subsidy would be much more obvious at the medical staff level. Perhaps hospitals would want to subsidize the premiums of their doctors with risky patients. Or such subsidies could be covered by a separate risk pool; this would probably require government intervention. The plan offers a major advantage over the present system in making the cross subsidies explicit.

Physicians often have privileges at several hospitals and deliver care in nonhospital settings as well. For the plan to work, physicians would have to select a hospital for purposes of purchasing medical malpractice insurance. Losses would be debited against the physician's hospital plan, regardless of where the loss occurred. Over time, hospital-based experience rating would encourage physicians to concentrate their hospital work at one hospital. Physicians would not want to be in groups with much higher (specialty-adjusted) expected losses than theirs unless each member could be made to pay a premium proportionate to his own expected loss. The internal allocation process may result in premium schedules not too different from those based on the individual physician's experience, but blatant errors in schedules would be less likely to occur because of better monitoring.

Other Groups. Large physician group practices in general, and Health Maintenance Organizations in particular, may also be experience-rating units. Peer review presumably takes place in such organizations.

HMOs now purchase medical malpractice insurance for their physicians and pay premiums based on the HMO's experience, or they self-insure. To the extent that HMOs' market share will grow, so will this form of group-based experience rating of medical malpractice premiums.

IV. Future Directions

Experience rating does make sense for medical malpractice. Pertinent statistical theory exists. Empirical evidence clearly indicates that past experience predicts future experience. Differences in expected loss within the risk-rating groups now used by physicians are large. With the increase in claims frequency, medical malpractice insurance is no longer a low-frequency line in several states. The limited amount of empirical evidence from other lines of insurance shows that insureds behave differently if their future premiums depend on their past actions. Although physicians undoubtedly face some uninsurable risk from medical malpractice, such as reputation loss and time spent in the litigation process, medical malpractice insurance is otherwise quite complete and must therefore blunt the deterrent effects of the tort system. Strengthening some forms of peer review would be desirable, and experience rating may be used to accomplish this objective.

Overall, malpractice insurers are lukewarm about experience rating. Evaluations of the success or failure of experience rating have not been based on in-depth analysis, but rather are based on impressions or practical experience.

Experience rating indeed does have some potential deficiencies—risk-classification error, the disincentive for physicians to report incidents at an early stage if the result of reporting is a higher premium subsequently, the disincentive for physicians to settle even if it is in the insurer's (and other policyholders) interest to do so, and the penalty on physicians who treat worthy, high-risk patients. Yet, for each of these, there are fixes, albeit imperfect ones. Risk-classification error may be reduced by using more accurate risk-classification categories not based on the individual physician's experience, or by rating at the level of the hospital medical staff or some other group. Incident reports can be excluded. A settled claim can be made to affect a premium less than a claim resolved at verdict. The premiums of physicians treating high-risk patients may be partly paid from a pool expressly established for this purpose.

Physician acceptance of experience rating is not high. This is so even though most physicians would benefit substantially from experience-rated premiums. There are several nonmutually exclusive explanations for this. First, physicians may not realize the extent of the cross subsidy. Second, they may fear mistakes in implementing experience rating. In particular, they may be charged for claims that they could not have prevented. Third, a solidarity principle may be at work. It may be desirable for the medical profession to keep the "bad apples" hidden. A more "selfish" explanation is that community rating provides insurance against becoming a bad apple.

There is, of course, considerable room for debate about plan design. This review suggests several features of a well-designed plan. Although claims frequency has been used almost exclusively in other lines, given the variation in severity and its positive correlation with frequency (Hassan and I), severity of losses should be used with frequency. Because of the long tail, it is appropriate to use open as well as closed claims. Because of risk-classification error, past experience of individual physicians should probably not be given full weight in premium setting. There is a role for peer review in combination with experience rating—peer review by small groups, such as hospital's medical staffs, to monitor quality of individual practitioners, by the insurer to determine the losses that should be counted for experience rating purposes, and by the state as part of the state's role in professional regulation to eliminate physicians with very bad claims records who are also bad doctors.

REFERENCES

Blumstein, James F. and Sloan, Frank A., "Antitrust and Hospital Peer Review," *Law and*

Contemporary Problems, Spring 1988, *51*, 8–92.

Ellis, Randall P., Gallup, Cynthia L. and McGuire, Thomas G., "Should Medical Professional Liability Insurance Be Experience Rated?," *Journal of Risk and Insurance*, forthcoming 1990.

Landes, Elizabeth M., "Insurance, Liability, and Accidents: A Theoretical and Empirical Investigation of the Effect of No-Fault Accidents," *Journal of Legal Studies*, April 1982, *25*, 49–65.

Lemaire, Jean, *Automobile Insurance: Actuarial Models*, Boston: Kluwer-Nijhoff, 1985.

______, "A Comparative Analysis of Most European and Japanese Bonus-malus Systems," *Journal of Risk and Insurance*, December 1988, *55*, 660–81.

Moore, Michael J. and Viscusi, W. Kip, "Promoting Safety through Workers' Compensation: The Efficacy and Net Wage Costs of Injury Insurance," *Rand Journal of Economics*, 1989, *20*.

Nye, Blaine F. and Hofflander, Alfred E., "Experience Rating in Medical Professional Liability Insurance," *Journal of Risk and Insurance*, March 1988, *55*, 150–57.

Palfrey, Thomas R. and Spatt, Chester S., "Repeated Insurance Contracts and Learning," *Rand Journal of Economics*, Autumn 1985, *16*, 356–67.

Rolph, John E., "Some Statistical Evidence on Merit Rating in Medical Malpractice Insurance," *Journal of Risk and Insurance*, June 1981, *48*, 247–60.

Sloan, Frank A., Bovbjerg, Randall R., and Githens, Penny B., *Insuring Medical Malpractice*, forthcoming 1990.

______ **and Hassan, Mahmud**, "Equity and Accuracy in Medical Malpractice Insurance Pricing," unpublished, 1989.

Sloan, Frank A., et al., "Medical Malpractice Experience of Physicians: Predictable or Haphazard?," *Journal of the American Medical Association*, December 1989, *262*, 3291–97.

Topel, Robert H., "Experience Rating of Unemployment Insurance and the Incidence of Unemployment," *Journal of Law and Economics*, April 1984, *27*, 61–90.

NEW CLASSICAL MACROECONOMICS IN THE OPEN ECONOMY[†]

International Transmission and Real Business Cycle Models

By ALAN C. STOCKMAN*

Recent progress on closed-economy models of "real business cycles" has inspired similar attempts at *quantitative* studies of business cycles in open economies. This work has several goals. One is to derive new implications of real business cycle models for variables that do not appear in closed-economy applications, such as the balance of trade and terms of trade. This can help us discriminate among alternative specifications of models and sources of shocks.

A second goal is to study the channels and dynamics of international transmission and to develop a more complete dynamic model of international trade. A third goal is to integrate two distinct but connected areas of research: business cycle models based on real disturbances, and models of exchange rates also based on real disturbances (see my 1980, 1987 papers). This will aid our understanding of exchange rates and provide new implications of the models of aggregate fluctuations and their international transmission.

The main features of aggregate business cycles in closed economies have been well documented. These features, such as procyclical investment and consumption, are *qualitatively* similar across countries.

There are also several key features of the international data that any model of aggregate fluctuations should explain: 1) Output is positively correlated across countries; 2) consumption is positively correlated across countries, but *less* correlated than output;[1] 3) Solow residuals are positively correlated across countries, but *less* so (or no more so) than output (Donna Costello, 1989); 4) the balance of trade surplus is countercyclical or acyclical (deficits are procyclical; see Backus and Kehoe, 1988).[2] These qualitative features of the data help limit the class of models and sources of exogenous disturbances worth considering. Additional limitations can be placed on models by the comovements of *prices*, such as the terms of trade and relative price of nontraded goods, with other aggregates.

I. One-Sector Models

Suppose we extend a "standard" real business cycle model based on productivity shocks (such as Finn Kydland and Edward Prescott, 1982, or Robert King et al., 1988), to a two-country setting, retaining *one* good in the model (Backus et al.; Marianne Baxter and Mario Crucini, 1989). This can provide interesting cross-country comparisons of the implications of the model, tying differences in business cycles to differences in parameters. This also highlights some critical features of the one-good model that make it difficult to explain the key observations mentioned in the introduction. Without going into details of this kind of model, certain features should be obvious (see Tesar, 1989).

[†] Discussants: Alessandra Casella, University of California-Berkeley; David Backus, Federal Reserve Bank of Minneapolis; Michael Mussa, University of Chicago.

*University of Rochester, Rochester, NY 14620. I am indebted to Donna Costello and Linda Tesar for helpful discussions, and Mike Pakko for help with programming. This research is supported by the National Science Foundation.

[1] This is related to the much discussed positive relation between national saving and investment (see Linda Tesar, 1990). David Backus et al. (1989) and Tesar present evidence on the cross-country correlations of consumption and output.

[2] Whether the trade balance is actually countercyclical or only acyclical is uncertain: the evidence depends somewhat on the sample and filtering method.

First, a temporary productivity shock in the home country creates a trade *surplus* in that country as domestic and foreign consumption rise. With repeated shocks, consumption tends to be *more* highly correlated across countries than is output. The key observations mentioned above contradict these two predictions.

Second, a permanent or persistent productivity shock in the home country can cause a trade deficit, because the shock raises the expected marginal product of investment. In equilibrium, investment in the home country can rise more than output, creating a trade deficit. But domestic and foreign consumption tend to *fall* in this case, while consumption is procyclical in the data.

Third, foreign and domestic consumption tend to move closely together in these models (with complete markets and no major distortions). To explain the lower cross-country correlation of consumption as compared with output, the model requires certain features. One possibility is that leisure and consumption are good substitutes in utility. Then a persistent home productivity shock raises the marginal product of labor and reduces leisure in the home country. If leisure and consumption are good substitutes, equilibrium consumption in the home country rises more than foreign consumption, or falls less; the close link between foreign and domestic consumption is broken.

But, fourth, a persistent or permanent productivity shock in the home country, by raising the expected marginal product of investment there, *reduces* foreign investment and foreign output. Physical capital moves from the foreign to the home country due to the higher expected return. Even if capital is not internationally mobile, foreign investment falls because the expected *value* of its marginal product falls (as output rises in the home country), and depreciation reduces the foreign capital stock. In either case, the foreign capital stock and foreign output fall. Foreign output may also be reduced by a fall in foreign employment, as the fall in capital reduces the marginal product of foreign labor.[3]

This kind of model could explain the positive cross-country correlation of output if productivity shocks were positively correlated. If that correlation is sufficiently high, it would dominate the endogenous tendency of outputs to move in opposite directions because of disinvestment in the foreign country. But Costello found that the cross-country correlation of Solow residuals (a standard measure of technology shocks) is *lower* (or certainly no higher) than the cross-country correlation of output. This suggests that the model misses important endogenous forces that tend to make foreign and domestic outputs move together (and make their correlation greater than that of technology shocks), and that tend to make foreign and domestic consumption less well correlated.

Finally, the model has *no* implication for the terms of trade, because there is only one good in the model. This is unfortunate because many models of exchange rates rely on real disturbances like those in real business cycle models. If real disturbances cause many of the large changes in exchange rates that we observe under floating-rate systems, then one-sector models of these disturbances miss some of their important effects.

The counterfactual *qualitative* implications of one-sector models do not imply that the real business cycle framework is useless for studying the international transmission of business cycles. In fact, models like this have had remarkable *quantitative* as well as qualitative success for the U.S. economy. Instead, the problems suggest that a multi-country real business cycle model must include features that are missing from existing closed-economy models. At the same time, these features must not interfere drastically with the success of those models for closed economies. Both theory and evidence on open economies suggest the inclusion of nontraded goods and multiple traded goods. Nontraded goods are quantitatively important and can play a role in breaking the link between foreign and domestic consumption. Terms-of-trade effects are essential for integrating the international business cycle model with models of exchange rates. Tesar and I are studying a model with these characteristics in ongoing research (also see Tesar, 1989). Rather than report detailed results here, I discuss some obvious *qualitative* changes in implications.

---

[3] If leisure and consumption are good substitutes in utility, foreign output could rise. If foreign consumption rises, it could lower the marginal utility of leisure and raise employment in the foreign country. If this effect dominated foreign disinvestment, foreign output could rise. It is unlikely that this effect is large enough to raise output. If it were, real wages in the foreign country would be countercyclical, which is inconsistent with real-wage behavior in the United States.

II. Qualitative Results of an Extended Model

Let us begin with a model (as in my paper with Harris Dellas, 1989) that adds nontraded goods to Robert Lucas's (1982) model. Add production to replace the endowments in that model. There are two countries and two traded goods with complete specialization in production, and also a nontraded good in each country. Production of each good is described by a Cobb-Douglas production function using labor and a type of capital specific to that good. A stochastic technology term multiplies each production function. Labor supply is fixed. Each good can be consumed or turned into capital specific to its own production in subsequent periods. There are four goods (two traded and two nontraded), so there are four capital stocks. Capital depreciates at a fixed rate.

A representative household in each country maximizes expected utility, with a standard utility function, over an infinite horizon. Nontraded and (both) traded goods are *complements* in consumption, in the sense that the cross derivatives in the utility function are positive.[4] As before, there are complete financial markets. This model can deliver implications for output, investment, consumption and savings, the trade balance, terms of trade, relative price of nontraded goods, and international transmission that are consistent with the key observations listed in the introduction: foreign and domestic output are positively correlated, and more so than foreign and domestic consumption or productivity shocks, and the trade balance is countercyclical.

A key feature of the model is the equilibrium response to productivity shocks in nontraded-good industries. Suppose there is a positive shock to productivity in the nontraded-good sector of the home country. Home consumption and output of the nontraded good rise. And home consumption of *traded* goods also rises, so home consumption rises more than home output. Foreign consumption of traded goods falls, so home and foreign consumption move in *opposite* directions in response to this shock. When shocks like this are repeated and combined with shocks to traded-good productivity, the cross-country correlation of consumption is positive but lower than that of output.

The fall in foreign consumption of traded goods raises the real interest rate in terms of these goods, and (through intertemporal substitution) raises foreign output. So output moves together in the two countries. Output is positively transmitted internationally, though consumption is negatively transmitted. The shock also causes a trade *deficit* in the home country, as in the data.

These models are capable of explaining the key observations mentioned in the introduction. But there is a nagging problem related to the behavior of relative prices. Models based on technology shocks tend to imply *countercyclical* relative prices of the goods subject to those shocks. But there is some evidence that Solow residuals are *positively* related to the relative price of a country's output.

Costello decomposed Solow residuals in two-digit industries in a sample of countries into components that are industry specific but common across nations, and components that are nation specific but common across industries. In my 1990 paper, I report the relations between Costello's series on the

---

[4] Traded and nontraded goods are often consumed as a bundle, as when nontraded retail services are a large fraction of the retail price of a "traded" good. This is also consistent with the evidence that traded goods (as well as nontraded goods) are more expensive in wealthier countries than in poorer countries.

nation-specific components of the Solow residuals and relative prices. The nation-specific component of the Solow residual (relative to another country) tends to be *positively* related to the relative price of the first country's consumption bundle. In other words, when the U.S. experiences a positive nationwide Solow residual relative to Canada, the relative price of the U.S. consumption bundle (measured by the exchange rate adjusted ratio of consumer price indexes) rises. The simple correlation for the Canadian-U.S. case from 1966–85 is .33; the correlation for the German-U.S. case from 1966–85 is .24. Eliminating the oil-price years of 1973–74, where there was a clear supply shock, raises the correlation in the German-U.S. case to .33 (and leaves the Canada-U.S. correlation roughly unchanged at .30).

This result is consistent with "demand" shocks such as disturbances to household production functions or tastes.[5] There is another reason to think demand shocks are important. Equilibrium models of exchange rates can more easily account for the high correlation between real and nominal exchange rates when there are disturbances to tastes or fiscal policies rather than technology; Dellas and I argued that this is evidence in favor of demand shocks.

The main issue involves the fact that increases in output tend to raise the demand for money. This creates a tendency for countercyclical nominal prices in response to technology shocks in real business cycle models, and there is evidence supporting this implication. But the countercyclical behavior of nominal prices is smaller than that suggested by standard estimates of money demand parameters. Moreover, the equilibrium exchange rate models explain the comovement of real and nominal exchange rates by deriving the change in the nominal exchange rate required to accomplish the equilibrium change in the terms of trade following a real disturbance. A positive technology shock in the home country lowers the relative price of the home country's traded good (in terms of the foreign country's traded good): this could be accomplished through depreciation of the domestic currency. Then the real and nominal exchange rates would move together. But the increase in money demand lowers domestic nominal prices, and tends to cause nominal appreciation of domestic currency. Only if this "money demand effect" is smaller than the aforementioned "relative price effect" will the real and nominal exchange rates move together as in the data. This requires that the inverse of the elasticity of substitution in consumption between goods be large relative to the income elasticity of the demand for money. The preceding argument applies to the effects of technology shocks. Demand shocks, on the other hand, can rationalize comovements in real and nominal exchange rates without these conditions, because they involve changes in relative prices but may not directly affect money demand (see my paper with Dellas).

These considerations suggest that Solow residuals do not represent pure technology shocks, but a conglomeration of shocks. If so, there are missing factors in the production function, misspecification, or mismeasurement of factor use with a countercyclical measurement error (as would occur, for example, with labor hoarding).

Disturbances to preferences (with a country-specific component) have the qualitative ability to rationalize the key observations cited earlier as well as the positive relation between Solow residuals and the relative price of the domestic country's consumption bundle. Consider a disturbance to preferences in the home country that temporarily raises the marginal utility of consumption of *all* goods. This causes a trade deficit, raises traded-good output in both countries, and raises nontraded-good output in the home (but not foreign) country. Consumption of traded goods rises in the home country and falls in the foreign county. The cross-country correlation of output, with repeated shocks like this, exceeds the cross-country correlation of consumption.

The taste shock also raises the relative price of the domestic consumption bundle, since it includes home nontraded goods. If the increase in home output shows up as a Solow residual in the production function (because of some misspecification similar to that caused by labor hoarding), the nation-specific component of the Solow residual will be positively related to the relative price. Similar effects can arise from taste shocks that shift demand from the output of one country to that of another; these shocks can also generate changes in real exchange rates.

---

[5] Define a demand shock as any shock except a shock to production functions or the system of property rights: examples are taste shocks, demographic changes, and fiscal policy changes.

III. Conclusions

There are many *qualitative* issues of model formulation that need further study. Theory and evidence from international trade suggest modifying the basic real business cycle model when applying it open economies. Reasonable specifications can be made qualitatively consistent with a set of observations about international comovements of output, consumption, Solow residuals, the trade balance, and relative prices. Even with unobservable shocks, such as taste shocks, the *implications* of such a model can be checked against the data. Ongoing research will establish whether such models are quantitatively as well as qualitatively consistent with observations.

REFERENCES

Backus, David K. and Kehoe, Patrick J., "International Evidence on the Historical Properties of Business Cycles," FRBank of Minneapolis Working Paper No. 402, 1988.

______, ______, **and Kydland, Finn,** "International Borrowing and World Business Cycles," FRBank of Minneapolis Working Paper No. 426R, 1989.

Baxter, Marianne and Crucini, Mario, "Explaining Savings-Investment Correlations," University of Rochester, 1989.

Costello, Donna, "A Cross-Country, Cross-Industry Comparison of the Behavior of Solow Residuals," University of Rochester, 1989.

King, Robert G., Plosser, Charles I. and Rebelo, Sergio T., "Production, Growth, and Business Cycles, I: The Basic Neoclassical Model," *Journal of Monetary Economics*, March/May 1988, *21*, 195–232.

Kydland, Finn E. and Prescott, Edward C., "Time-to-Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Lucas, Robert E., Jr., "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, November 1982, *10*, 335–60.

Stockman, Alan C., "A Theory of Exchange Rate Determination," *Journal of Political Economy*, August 1980, *88*, 673–98.

______, "The Equilibrium Approach to Exchange Rates," *Economic Review*, FRBank of Richmond, March-April 1987, *73*, 12–31.

______, "Solow Residuals and Real Exchange Rates," University of Rochester Working Paper, forthcoming 1990.

______ **and Dellas, Harris,** "International Portfolio Nondiversification and Exchange Rate Variability," *Journal of International Economics*, May 1989, *26*, 271–90.

Tesar, Linda, "Nontraded Goods, Risk-Sharing, and Trade in Capital," Working Paper, University of Rochester, 1989.

______, "Savings, Investment, and International Capital Flows," *Journal of International Economics*, forthcoming 1990.

Bargaining and International Policy Cooperation

By KENNETH ROGOFF*

The past decade has witnessed the growth of a large literature on international cooperation in trade and macroeconomic stabilization policy. Virtually all the models developed to date, however, are based on one of two extreme assumptions concerning governments' ability to commit to international agreements. Either they assume that governments can make constitutionally binding long-term agreements, or else they assume that governments have no ability to make legal commitments whatsoever. In the latter case, international policy cooperation is possible only to the extent that reputational factors will allow.[1]

In this paper, I consider a world in which there is no legal mechanism for enforcing long-term international agreements, but where governments must still incur some small direct costs if they renege. These small costs might arise due to legislative or administrative frictions. I also allow for the possibility that international economic policy agreements can include small sidepayments. For example, in negotiating a bilateral reduction in tariffs, two allies could simultaneously agree to redistribute the burdens of defense expenditures.

I. Bargaining and the Flow Gains from Trade

Allowing for small side payments and small costs of reneging can have a dramatic effect on short-term agreements involving *flows* (for example, the gains from *intra*temporal trade). Consider a simple two-country, two-good model in which the representative citizens of both countries have identical homothetic tastes. Production is exogenous and constant. There are gains from trade because the two goods are produced in different proportions across the two countries. Clearly, the laissez-faire free-trade equilibrium is Pareto efficient. It will not necessarily come about, however, unless the two governments can somehow agree on a mutual nonintervention pact. The problem is that if either government could consistently get away with imposing tariffs *unilaterally*, it would be able to make its citizens better off at the expense of those residing in the other country. Tariff warfare is a classic prisoner's dilemma problem.

Let us assume that the two countries' governments have finite horizons and that the Nash equilibrium to the tariff war is unique. In this case, the laissez-faire equilibrium cannot be supported by reputational factors. It can, however, be achieved if there are small legislative costs to breaking an agreement. Because gains from trade are a flow, neither country will have any short-run incentive to defect from a free-trade agreement. Moreover, once an agreement is broken, both countries will have every incentive to replace it with a new one. Note that free trade can be maintained here even in the last period.

If side payments are possible, efficient trades will still take place, but the gains from trade will be divided according to the relative bargaining power of the two countries. Suppose, for example, that country A enjoys much larger gains from trade than country B (perhaps because A is smaller and less diversified). In this case, country B may be able to extract a side payment from country A in return for agreeing to free trade. Note that B can only extract a small side payment at any given point in time, since it cannot commit to refrain from asking for more money in the

---

*Economics Department, University of California, Berkeley, CA 94720. This research has been supported by the NSF under grant SES-87-20800, and by the Alfred P. Sloan Foundation. I am grateful to Jeremy Bulow for many discussions on this topic.

[1] In some areas of economic policy, restraints on cooperation can actually improve welfare. My 1985 paper for the case of monetary policy, and Patrick Kehoe (1989) for the case of fiscal policy, provide examples where the ability to coordinate policies would exacerbate governments' credibility problems vis-à-vis their own private citizens.

future. But this restriction is unimportant here since the gains from trade are a flow, and *B* can extract a steady stream of payments. Of course, to gain insight into what factors govern the size and the direction of the sidepayments, it is necessary to have a formal model.

As of yet, there exist very few theoretical analyses based on the paradigm suggested here; one special case has been developed by Jeremy Bulow and myself (1988b). We base our analysis on an alternating offers bargaining framework,[2] in which a key feature of bargaining is that the parties can only strike bargains over current flows. Since any long-term deal is subject to renegotiation, countries must constantly recontract. In our setup, both countries are risk neutral and there is full information. Consequently, in equilibrium a deal is always reached without delay, and side payments depend only on the relative magnitude of each countries' gains from trade and on their relative discount factors. If the two countries were risk averse, however, the richer country might be able to bargain for a larger share of the gains from trade. With risk aversion, the poorer country has more to lose from a breakdown in trade. Thus the bargaining paradigm suggests a loose sense in which a free-trade equilibrium can involve strategic exploitation.

Once one departs from a pure endowment economy and allows for investment, the ability to make bargains over flows is no longer enough to achieve an efficient equilibrium. The problem is that countries will have an incentive to distort investment patterns to improve their bargaining position in future periods.

Clearly, a bargaining paradigm only makes sense in situations where individual countries have monopoly power. If the world consists of a large number of countries and the usual assumptions for perfect competition are met (for example, no increasing returns to scale), then of course strategic problems disappear. For example, if a country can import its wheat from any of a large number of competing countries, then there cannot be any bargaining incentive for it to subsidize an uncompetitive wheat-farming sector.

II. Bargaining and Mercantilism

History provides an important example of how bargaining factors can dictate trade policy. Modern textbooks almost universally condemn the mercantilists for arguing that countries should always strive to run trade surpluses. After all, a government's objective ought to be to maximize its citizens' utility, not their wealth. But as Eli Hecksher (1955) noted, the mercantilists had completely rational reasons for their policy recommendations. Back in mercantilist times, countries frequently had to rely on mercenary armies to fend off invaders or to protect colonies. To ensure survival, it made sense to have an ample supply of gold on hand. Even during peacetime, a country's ability to project military power improved its bargaining position vis-à-vis its colonies and other competing empires.

III. Reputational Mechanisms for Enforcing Trade Agreements

Standard reputational models provide a different approach to analyzing international policy cooperation (see Matthew Canzoneri and Dale Henderson, 1988). A prototypical example arises in the case where governments have infinite horizons and do not discount the future very heavily. Then the laissez-faire equilibrium can be maintained if both countries believe that any defection from free trade will ultimately lead to a prolonged state of tariff warfare. If governments have high discount rates, then some cooperation is still feasible, but not enough to achieve efficiency. How realistic is the repeated-game reputation paradigm? One can plausibly argue that the postwar consensus to adhere to free trade was strengthened by memories of the tariff wars of the 1930s. On the other hand, a problem with the reputational models is that they admit a broad range of equilibria, and a substantial degree of coordination seems to be required to achieve the optimal one (see my 1989 paper). Also, the assumptions underlying the present bargaining approach appear to be more realistic in many contexts.

---

[2] The analysis extends Ariel Rubinstein's 1982 model to the case where two parties continually bargain over flow gains from trade, and where any long-term agreement is subject to recontracting.

It is not my intention here to criticize earlier approaches to analyzing cooperation, since these approaches have yielded some important insights. Rather, I hope to show that a bargaining-theoretic view of policy coordination provides a fresh and interesting perspective on a variety of problems.

IV. Bargaining and Net Capital Flows

Small legislative frictions and small lump sum side payments are of far less use in striking deals over *inter*temporal trade. When the costs to reneging are small, countries have no way to commit to making large lump sum payments; this precludes efficient allocation of international investment funds as well as efficient risk sharing. Reputational factors can in principle support significant net loans across large countries, but there are a number of reasons (such as coordination problems) for believing that these are not always operative.[3]

Thus our bargaining model suggests that net international lending may be much smaller than a general equilibrium model might predict. Indeed, to a first approximation, this view may be accurate. Net capital flows have seldom been very large historically; this observation is the mirror image of Martin Feldstein and Charles Horioka's (1980) finding that savings and investment are highly correlated across countries. Clearly they must be correlated if countries are unable to commit to making large net transfers. Many other possible explanations of the Feldstein-Horioka finding have been advanced, but the above rationale is perhaps the simplest. It is also potentially consistent with the observation that *gross* international capital flows can be quite large. Bargaining problems are likely to be most severe when a country is a large *net* debtor or creditor.

Obviously, the above discussion abstracts from some subtle and important domestic distribution questions. Among home residents, those who hold assets abroad are likely to have very different interests from those who are net debtors to foreigners.[4] It is also important to recognize that a country may have a positive net asset position in some countries that is offset by negative positions in others. Understanding how heterogeneity affects the scope for intertemporal trade is an important topic for future research. Nevertheless, a simple bargaining framework may well provide a more realistic picture of intertemporal international trade than the classical complete markets macro model does.

V. Bargaining and LDC Debt Plans

A bargaining approach also provides an interesting perspective on plans for alleviating the Third-World debt crisis. For example, one important class of plans involves the creation of a new multilateral lending organization (see Peter Kenen, 1983). The new agency would issue bonds and use the proceeds to buy up Third World government debt at discount. It would then pass on the discounts to troubled Third World debtors, thereby helping to relieve the debt overhang problem. This is an interesting and important idea, but there are two possible drawbacks. The first is that it would not be easy for the new debt agency to buy up the debts without bidding up their prices (see Bulow and myself, 1988a, 1990). I will ignore this problem here. The second has to do with the implicit assumption that once developing-country debts have passed into official hands, they will be paid off like clockwork. A bargaining approach casts doubt on this proposition. Historically, developing-country governments have been successful in extracting positive net resource flows from industrialized-country governments. (Private creditors are much tougher bargainers.) Why should the creation of some paper claims have any fundamental effect on this process? Bulow and I (1990) argue that developing countries would never make significant repayments to a multilateral debt agency, and that industrialized-country taxpayers would ultimately be forced to honor the agency's own bond debts.

---

[3] Bulow and I (1989) argue that, in theory, it is very difficult for small countries to have a reputation for repayment. If so, then a small country must try to borrow under the umbrella of a large country's legal system.

[4] Bulow and I (1990) emphasize that the bargaining factors underlying private international debt negotiations are often quite different than those underlying debt negotiations between governments.

VI. Conclusions

To date, there has been only a limited amount of formal research applying bargaining theory to international policy cooperation. I have argued here that this paradigm is a fruitful one for analyzing a broad range of problems. Under certain conditions, bargaining can lead to efficient intratemporal trade. However, the gains from trade are divided according to the relative bargaining power of the two countries and not simply according to the difference between autarky and free-trade utility levels.

In general, achieving efficient intertemporal trade is much more problematic. Bargaining problems may explain why net international capital flows are relatively small (the Feldstein-Horioka hypothesis) whereas gross flows are much larger.

REFERENCES

Bulow, Jeremy and Rogoff, Kenneth, (1988a) "The Buyback Boondoggle," *Brookings Papers on Economic Activity*, 2:1988, 675–98.

______ **and** ______, (1988b) "Multilateral Negotiations for Rescheduling Developing-Country Debt," *International Monetary Fund Staff Papers*, December 1988, *35*, 644–57.

______ **and** ______, "Sovereign Debt: Is to Forgive to Forget?," *American Economic Review*, March 1989, *79*, 43–50.

______ **and** ______, "Cleaning Up Third-World Debt Without Getting Taken to the Cleaners," *Journal of Economic Perspectives*, Winter 1990, *4*, forthcoming.

Canzoneri, Matthew and Henderson, Dale W., "Noncooperative Monetary Policies in Interdependent Economies," unpublished, 1988.

Feldstein, Martin and Horioka, Charles, "Domestic Saving and International Capital Flows," *Economic Journal*, June 1980, *90*, 314–29.

Hecksher, Eli, *Mercantilism*, London: Allen and Unwin, 1955.

Kehoe, Patrick J., "Policy Cooperation Among Benevolent Governments May Be Undesirable," *Review of Economic Studies*, April 1989, *56*, 289–96.

Kenen, Peter B., "A Bailout Plan for the Banks," *New York Times*, March 6, 1983.

Rogoff, Kenneth, "Can International Monetary Cooperation Be Counterproductive?," *Journal of International Economics*, May 1985, *18*, 199–217.

______, "Reputation, Coordination and Monetary Policy," in Robert Barro, ed., *Handbook of Modern Business Cycle Theory*, Cambridge: Harvard University Press, 1989.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, *50*, 97–109.

The New Classical Macroeconomics and Stabilization Policy

By RUDIGER DORNBUSCH*

The new classical macroeconomics (NCM) is not particularly useful in explaining stark episodes. To be successful, the new classical approach will have to be developed in a number of areas including coordination problems and political economy issues of stabilization. These modifications can in principle be worked into the NCM classical approach, but that will change the main thrust. So far, the message reads: markets clear, productivity shocks are the dominant source of fluctuations, intertemporal substitution in consumption and labor supply are the chief amplifying factors, and government policy is substantially irrelevant. On these key propositions, the evidence fails and the emphasis is misdirected. The issue is not whether markets work, but rather how they work, a concern rightly emphasized by the new Keynesian economics.

I. New Classical Economics

The NCM models the maximization problem of the individual rational and forward-looking agent.

Three implications follow from the methodology and are the distinguishing characteristic of this approach. The role of relative prices is pervasive once the proper distinctions are made between intra- and intertemporal terms of trade, and between price changes that are current and future, transitory and permanent, and anticipated and unanticipated. Next, because economic agents' policy environment is part of the information set and budget constraint, we can obtain informed answers to questions about the effects of policy realizations or changes in rules. Finally, the derivation of aggregate equations from basics constitutes the basis for "deep" as opposed to "opportunistic" or "*ad hoc*" analysis, including informed welfare judgments.

New classical economics in the open economy has no comprehensive statement. But a set of results and presumptions helps characterize what it is all about:

Rigorous insistence that markets work; individual self-interest assures that relative prices adjust to assure full resource utilization. Persistent involuntary unemployment is ruled out except when an explicit mechanism bars agents from pricing themselves into jobs. Variations in economic activity are not excluded by the approach, but they represent optimal adjustments of labor effort to the time path of productivity and real relative prices.

Governments optimally select spending and financing policy (see Robert Lucas, 1986). The tax rate is therefore intertemporally flat, and seigniorage is integrated in the optimal tax structure. Government spending is motivated by the optimal provision of public goods.

In the absence of capital controls, international capital markets are tightly linked so as to equalize *ex ante* rates of return adjusted for anticipated exchange rate changes and diversification considerations.

Nondistortionary tax changes are entirely neutral. Even in a Ricardian setting, government spending changes have sectoral demand effects (unless they are fully valued by households) and therefore affect relative prices over time and in the steady state. Because they affect relative prices, government spending changes do affect the current account whenever a distinction is made between traded and nontraded goods or between importables and exportables.

Monetary policy is entirely ineffective. Changes in the nominal money stock affect only prices and the exchange rate or translate into offsetting reserve changes. Real variables, specifically the real exchange rate,

---

*Ford International Professor of Economics, MIT, Cambridge, MA 01239, and Research Associate, NBER. Financial support was provided by a grant from the National Science Foundation.

TABLE 1—MEXICO: ECONOMIC INDICATORS
(Index 1980 = 100)

	1981	1983	1985	1987
TOT	97	78	72	57
Y	108	102	108	106
YPC	105	94	96	91
Manufacturing:				
Y	106	96	106	105
W	103	75	71	61
E	105	91	92	85

Source: Banco de Mexico.
Note: Y = real GDP, YPC = per capita real GDP, TOT = terms of trade; W = real product wage in manufacturing, E = employment in manufacturing.

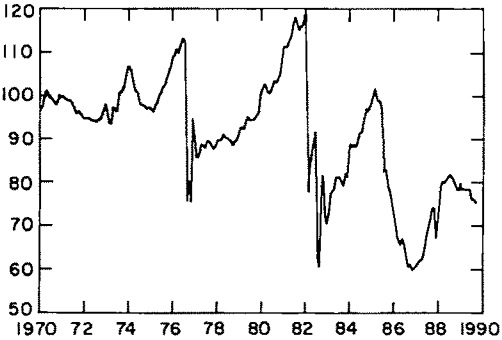


FIGURE 1. THE MEXICAN REAL EXCHANGE RATE
(Index 1980–82 = 100)

the real interest rate, or economic activity would be unaffected. Changes in the growth rate of money, with offsetting neutral reduction in taxation, would likewise have no real effects except to the extent that portfolio composition is shifted toward or away from external assets.

The behavior of the real exchange rate ought to be independent of the nominal exchange regime. Specifically, a shift from a fixed to a flexible exchange rate, or to a crawling peg should leave the real exchange rate unchanged in terms of its level or variability. A nominal devaluation does not imply a real devaluation.

II. The Mexican Experience

A strong test of a theory is to expose it to a stark and discriminating body of facts. The NCM predictions have been tested extensively in regard to the relationship between budget deficits, debt, and consumption and the cyclical correlations properties of macro series. Few studies offer support for the NCM over more opportunistic models. Imaginative work on exchange rate regimes comes to the same conclusion (see Michael Mussa, 1986, and Alan Stockman, 1987).

A simple case study of Mexico (see Table 1) points in the same direction. How does the NCM help explain the experience of Mexico? Per capita income and employment fell significantly and the real wage declined by nearly 40 percent. A starting point for any approach is the recognition that the debt crisis of 1982, and the subsequent credit rationing, must be important influences as should be the behavior of the real oil price.

In respect to the external balance, the NCM does not have a strong prediction—nondistortionary tax changes have no effect, but transitory government spending changes do affect the external balance. A regression with annual data for the period 1965–87 shows that the noninterest external balance ($NICA$) includes the primary budget (BUD), the log of the real price of oil (Oil), and the real exchange rate (REX) as significant explanatory variables.[1] An increase in the budget deficit by 1 percent of GNP worsens the external balance by a quarter of a percent. With budget changes reflecting at least in part changes in government spending, these results do not discriminate between the NCM and alternatives.

$$(1)\quad NICA = \underset{(-6.2)}{-68.0}\ \underset{(-2.1)}{-0.25BUD} \;\; + \underset{(3.2)}{2.24Oil} - \underset{(5.8)}{14.1REX} \qquad R^2 = 0.90$$

Consider next real exchange rates. Figure 1 shows the real exchange rate measured as Mexican relative to competitors' manufactures prices expressed in a common currency. The NCM predicts independence of the real exchange rate (market-clearing relative prices) from the nominal exchange rate regime. That is not the case: the maxi devaluations of 1976, 1982, or 1987 show clearly in the data. Unlike in the case of floating rates, the timing of devaluations was discretionary and followed with a significant lag the accumulation of imbalances. Thus it cannot be argued that they reflect the timing of changes in equilibrium relative prices. This constitutes strong evidence against the NCM.

---

[1] $NICA$ and BUD are measured as a percent of GNP and the real oil price and the real exchange rate are expressed as logarithms. The equation was corrected for second-order serial correlation.

A strong case against the NCM comes also from the behavior of real wages and employment. The classical model does not help explain the large decline in real wages and employment in Mexico—just as it fails to explain the same facts in the Great Depression. A positive correlation of real wages and employment requires labor demand shocks. Suppose labor demand derives from a competitive industry with a CES production function. Using the observed data on output, employment, and real wages, this model implies that there must have been a *favorable* productivity shock of at least 20 percent between 1980 and 1987. Few observers would seriously assert that Mexican macroeconomic performance in the 1980s reflects a favorable shock.

Reality also departs from the new classical model in the area of government policy. The Mexican experience reflects U-turns in fiscal policy that are simply ruled out by the forward-looking, intertemporally optimizing model of government. In the 1970s, prior to the debt crisis, increased oil revenues translated into a boom in *current* spending by the government. In the 1980s, budget cutting became the rule. New classical theory would have predicted that, in response to the interest rate shock and oil price decline, the government would have responded with a permanent increase in tax rates; in fact, the government reduced subsidies and sharply cut back public sector investment and social spending. Moreover, stabilization has not been carried far enough to renew investment and growth.

I have listed a few inconsistencies and a careful study might harvest more. A model that explains the experience of Mexico or other Latin American economies today needs special features in several areas. One is that the treatment of investment and capital repatriation must be linked to the value of waiting, a point developed below. Another is that a theory of economic policy must explain why government stabilization is not carried to the point where the private sector carries on with substantial confidence.[2]

III. Capital Repatriation and the Waiting Option

A common problem in the aftermath of stabilization is the lack of a stabilizing capital reflow. Moreover, when capital does return, it is placed in liquid form in financial markets rather than in plant and equipment. Investors have an option to postpone the return of flight capital. They will wait until the front loading of investment returns is sufficient to compensate them for the risk of relinquishing the liquidity option of a wait-and-see position. There is definitely little commitment to a rapid resumption of real investment—the reason being the residual uncertainty over whether stabilization can in fact be sustained.

I concentrate on a simple formulation that assumes that the economy (Mexico) has two states. In the good state, the return on an investment is r^g, and r^b in a bad state. Investors have the option to invest abroad (in Miami) at r^* or at any time to make an irreversible investment in Mexico. Their evaluation of states follows a Markov process: In a bad state there is a probability q of persistence and $1-q$ of a shift to a favorable state. As a sharp simplification, once a favorable state prevails, it is expected to last forever. Investors are assumed risk neutral.

How much of a front-end premium, Φ, over the Miami return is required for an investor to go ahead and invest rather than wait and see, maintaining the option of entering once the favorable state is verified.

---

[2] See my 1989 working paper for a positive theory of stabilization.

The required front-end premium is

$$\Phi_1 = [q/(R^* - q)](r^* - r^b), \tag{2}$$

where $R^* = 1 + r^*$. This formulation has two key features. First, it confirms Ben Bernanke's (1983) Bad News Principle that the option value of waiting depends only on the bad news, not the good news. The reason is that investors can avail themselves of good news situations by investing even late. Second, if bad states are very persistent, the premium approaches the present value of the differential, $(r^* - r^b)/r^*$. Thus persistence translates into a sizable front-end premia required to bring about immediate commitment.

The assumption that latecomers can obtain the same return as those who invested early is perhaps unrealistic. Suppose now that latecomers obtain in the year of investment only a fraction α of the good return, but thereafter participate equally. The formula for the front-end premium then becomes

$$\Phi_2 = \Phi_1 - [(1-q)/(R^* - q)](1-\alpha)r^g. \tag{3}$$

The penalty for latecomers thus naturally reduces the front-end premium. But it is apparent that if bad states are expected to be highly persistent, this latter effect becomes small.

These ideas can be carried a step further if we assume that there is a link between the front-end premium and the extent of capital reflow. Such a relation can exist either because a reflow reduces the probability of a bad state, or it raises returns in unfavorable states. We assume then that $\Phi = \Phi(K, \ldots)$ with a larger capital inflow reducing the premium (i.e., $\Phi'(K) < 0$).

The return on assets in Mexico, m, is taken to be exogenous to the reflow. The criterion for the excess return in Mexico required to induce repatriation now becomes $m > \Phi(K)$. Figure 2 shows the schedule representing the return on assets in Mexico, m, and the required return $\Phi(K)$. There are two possible equilibria. If no capital repatriation is anticipated, then individual asset holders perceive a required return $\Phi(0)$ and see an actual return of m that is insufficient to cover their risk. As a result they do not repatriate, and hence the bad equilibrium prevails. Because no capital has returned, the equilibrium may be self-fulfilling in its assumption of a high probability of program failure.

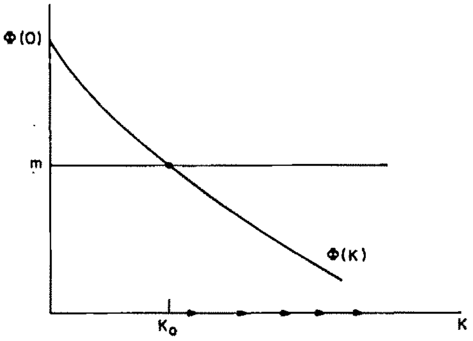


FIGURE 2. THE CAPITAL REPATRIATION PROBLEM

In the other equilibrium, the individual investor expects at least K_0 to be repatriated. With so much repatriation, the risk assessment drops sharply and every investor repatriates. As a result, *all* capital comes back because the required return is below the actual payoff. The question then is how to trigger this "good" equilibrium.

How can governments reassure investors? The common answer is to bring about a "credible" stabilization. If real depreciation is not sufficient to bring about investment, the government faces a very awkward position: income is being redistributed from labor to capital, but because the real depreciation is not sufficient, the increased profits are taken out as capital flight. Labor will obviously insist that the policy be reversed. This uncertainty is an important feature in understanding the real exchange rate—capital flight relationships and the poststabilization difficulties. The options to postpone repatriation and to postpone investment in plant and equipment, in export markets or simply in working capital, are too valuable, and hence growth does not return.

The discussion of the option value of waiting, and the associated credibility issue, highlights one way in which the competitive model of the NCM fails to address the transition from stabilization to growth. Stabilization by itself is not enough to trigger a virtuous circle. There is a need for a coordination mechanism that overcomes the competitive market tendency to wait.

The point can be taken a step further to bring in political economy considerations. There are economic equilibria and there are political ones. Open-economy issues must be modeled with both in mind. An extraordinarily large adjustment in real wages may set the economic incentives right, but it may bring about a political situation that is not comforting for investors. Similarly, the direction and even size of required economic adjustments are understood, but politically these are not possible. What then?

What markets consider a sufficient policy action may simply be beyond the political scope of democratic governments. In fact, if governments went far enough to create the incentives that would motivate a return of capital and the resumption of investment on an exclusive economic calculation, the implied size of real wage cuts might be so extreme that now, on political grounds, asset holders might consider the country too perilous a location. In the aftermath of a major macroeconomic shock, competitive markets by themselves may be unable to restore a good equilibrium.

REFERENCES

Bernanke, Ben, "Irreversibility, Uncertainty and Cyclical Investment," *Quarterly Journal of Economics*, February 1983, *98*, 85–106.

Dornbusch, Rudiger, "Credibility and Stabilization," Working Paper, MIT, 1989.

Lucas, Robert, "Principles of Fiscal and Monetary Policy," *Journal of Monetary Economics*, March 1986, *17*, 117–34.

Mussa, Michael, "Nominal Exchange Rate Regimes and the Behavior of Real Exchange Rates: Evidence and Implications," *Carnegie-Rochester Conference Series on Public Policy: Real Business Cycles, Real Exchange Rates and Actual Policies*, 1986, *25*, 117–214.

Stockman, Alan, "The Equilibrium Approach to Exchange Rates," *Federal Reserve Bank of Richmond Economic Review*, March/April 1987, *16*, 12–30.

Equilibrium Short Horizons of Investors and Firms

By ANDREI SHLEIFER AND ROBERT W. VISHNY*

This paper attempts to explain the often lamented pursuit by investors of short-term capital gains and the selection by firms of short-term investment projects. Our starting point is the observation that, in practice, arbitrage (trading based on knowledge that the price of an asset is different from its fundamental value) is cheaper for assets that cannot stay mispriced for long (short-term assets) than for assets that can (long-term assets). In equilibrium, the net expected return from arbitrage in each asset must be the same. Since arbitrage in long-term assets is more expensive than it is in short-term assets, the former must be more mispriced in equilibrium for net returns to be equal. Thus the rational behavior of arbitrageurs leads to greater mispricing of long-term assets in equilibrium.

Moreover, managers of firms are typically averse to severe underpricing of their equity because they risk getting fired or taken over. They should then avoid investments that raise the cost of arbitrage of their equity. Since mispricing of claims to long-term investment projects can take a long time to disappear, such projects will be avoided. In this way, short horizons of arbitrageurs lead to short horizons of corporate managers.

We stress the distinction between long (short) term investment projects and long (short) term assets. Long (short) term *investment projects* have distant (proximate) cash flows. However, an arbitrageur cares not when an investment project pays off, but rather when the mispricing of the claim to this project disappears, so that he can reap the rewards from his trade. A short-term *asset* is one where mispricing must disappear in the near future, whereas mispricing of a long-term *asset* can persist for a long time. The time to disappearance of mispricing depends on how fast fundamental uncertainty about an asset is resolved, how fast investor misperceptions are corrected, and how rapidly arbitrage drives the price to fundamental value. In general, mispricing of claims to long-term investment projects takes longer to disappear since it takes longer for fundamental uncertainty to be resolved. As a result, long-term assets are typically claims to long-term investment projects. Such association, however, is not automatic.

Examples of short-term assets are options, futures, and other instruments that have a fixed and relatively short expiration time. Examples of long-term assets are stocks and foreign exchange, where mispricing can take a long time to correct. Sometimes even stocks can be short-term assets. For example, an arbitrageur betting on the outcome of a takeover bid, or on an imminent earnings or other public announcement, can expect the mispricing that he is betting against to disappear quite fast. Such an arbitrageur can liquidate his position once the takeover bid fails even if he believes that the stock is still underpriced, and does not need to wait until *that* mispricing is corrected. What matters to the arbitrageur, then, is the horizon of the disappearance of mispricing

In Section I we argue that arbitrage in short-term assets is much cheaper than arbitrage in long-term assets. Section II builds this observation into a model in which long-term assets must be more mispriced in equilibrium than short-term assets. Section III shows how short horizons of smart investors

---

[†] *Discussants*: Glenn Hubbard, Columbia University; Herbert Gintis, University of Massachusetts-Amherst; Michael Woodford, University of Chicago.

*University of Chicago, Chicago, IL 60637.

translate into short horizons of firm managers averse to underpricing of their equity (see also Jeremy Stein, 1988; 1989). Section IV shows how the flight of smart money to short-term arbitrage can be self-fulfilling.

I. The Relative Costs of Long- and Short-Term Arbitrage

By greater relative cost of long-term arbitrage, we mean that the expected carrying cost of a $1 investment held until the mispricing is eliminated is higher for the long-term arbitrage. This is not just a reflection of the time value of money. An arbitrageur with access to a perfect capital market does not care how long it takes a mispriced security to reach its fundamental price. Suppose that an asset with a fundamental value of $6 trades today for $5. Let the interest rate be 10 percent, which is also the rate of return on the fundamental value of the asset and the arbitrageur's cost of funds in a perfect capital market. In this case, the arbitrageur does not care when the price of the asset catches up with its equilibrium value. If it does so next period, the arbitrageur sells the asset for $6.60, repays $5.50 to his creditors, and gets $1.10 which has the present value of $1. If the price reaches fundamentals two periods from now, the arbitrageur sells the asset for $7.26, repays $6.05 to his creditors, and again ends up with $1 in present value. Moreover, if the price of the asset does not reach its fundamental value early, and the arbitrageur discovers another opportunity in the meantime, he can borrow more and pursue that one as well. There is no opportunity cost of tying up one's funds. The same irrelevance argument holds for short sales, and for the case of uncertainty where the arbitrageur can costlessly issue equity to sell off his risk.

Arbitrage, however, is often risky and risk cannot be completely sold off in the market. If, for example, an asset is underpriced relative to its fundamental value, and a smart investor buys it, he has to bear the risk that before mispricing is eliminated or reduced the fundamental value actually falls. In this case, his arbitrage trade results in a loss even though it was *ex ante* attractive. In addition to fundamental risk, the smart investor bears the risk that the mispricing gets worse before it is eliminated, called "noise trader risk" by Bradford De Long et al. (1990). If the smart investor has for some reason to liquidate his position when mispricing gets worse, his arbitrage trade results in a loss. Both fundamental and noise trader risk are more important for assets where the elimination of underpricing takes longer, since there is more time for bad news or a wave of pessimism to hit. These risks raise the cost of arbitraging long-term assets relative to short-term assets.

These risks can be perfectly shared in a perfect capital market that is well informed about the arbitrageur's true ability. But in practice, outside investors do not know whether an arbitrageur is smart, and worry that he might take risks with their money without earning an extra return. As a result, they restrict the supply of funds to the arbitrageur (Joseph Stiglitz and Andrew Weiss, 1981). First, his cost of funds exceeds the (risk-adjusted) interest rate at which he would be able to borrow if information was symmetric. Second, he faces a limit on the amount he can borrow. If in our earlier example the return on the fundamental value is 10 percent, but the cost of funds to the arbitrageur is 12 percent, he obviously prefers that mispricing be eliminated sooner.

Credit constraints impose an additional cost of long-term arbitrage relative to short-term arbitrage, namely the opportunity cost of having one's money tied up. When a smart investor dedicates the limited resources that he can get from others to an arbitrage position, he obviously cannot dedicate them to another position. The longer it takes for the price of a mispriced asset to return to its fundamental value, the longer is the period during which the arbitrageur's resources are tied up, and therefore the higher is the opportunity cost of this arbitrage trade.

Since the high cost of funds and the credit constraint reduce the arbitrageur's profits, he will try to convince his lenders that he is indeed smart and can earn abnormal returns, so that they lend him more and at more attractive rates. To do that, he will try to show good performance repeatedly and *fast*, presumably by making multiple short-term arbitrage trades rather than a single long-term trade. Demonstrating one's talents with long-term trades is expensive and risky by comparison. In this way, not only do the capital market imperfections raise an arbitrageur's cost of funds and so lead to a preference for short-term arbitrage, but the arbitrageur's efforts to reduce these problems and the cost of funds lead to a further bias for short terming.

As feared by Keynes, smart money does not flock to long-term arbitrage: it requires a lot of patience, and patience is costly.

II. A Simple Model

We assume there are three periods: 0, 1, and 2, and two types of investment projects: long and short term. Each firm undertakes a project of one of the two types. Both types of projects have a cash flow only in period 2. The expected period-2 payoff from a short-term project is V_s and the expected payoff from a long-term project is V_l. The true value of a short-term project becomes known in period 1, at which point its mispricing is eliminated. The same happens with a long-term project in period 2. This specification makes clear that arbitrageurs care not when the cash flow of a project is, but when mispricing disappears. All investors are risk neutral. We assume that the market interest rate is 0; the expected payoff from a project is then equal to its fair market value.

There are two types of investors in the model: noise traders and smart money. In period 0, noise traders are either pessimistic or optimistic about the value of each project i (whether long or short term). When noise traders are pessimistic, their combined period-0 demand schedule for the equity of firm i is given by

$$(1)\qquad q(NOISE,i)=[V_i-S_i]/P_i,$$

where $S_i>0$ is the pessimism shock and P_i is the price. When noise traders are optimistic, the demand curve is the same, except S_i is added rather than subtracted. Whether noise traders are bullish or bearish is known in period 0. We only consider the bearish case; the other case is symmetric. We assume that each stock i is in unit supply. Without a shock to their beliefs, noise traders would hold the whole supply at the fundamental price and there would be no need for arbitrage.

We assume that arbitrageurs have no money of their own, but each can borrow in period 0 up to the amount b at the gross interest rate $R>1$. The cost of funds to the arbitrageurs in excess of the riskless rate of 0 and the credit constraint reflect the imperfections in the loan market, which we do not model explicitly. In this model, each arbitrageur actually wants to borrow all of b.[1] Since agents are risk neutral, they need not diversify, and we can assume that each arbitrageur invests all of b in the same asset i. If n_i arbitrageurs invest in asset i, their combined demand curve is

$$(2)\qquad q(SMART,i)=n_i b/P_i.$$

Because of the borrowing constraint, arbitrageurs also have unit elastic demand. In period 0, demands for each asset must add up to the unit supply, so,

$$(3)\qquad P_i=V_i-S_i+n_i b.$$

We assume that the total number of arbitrageurs is not sufficient to bring prices of all assets to fundamentals in period 0: $n_i b<S_i$. In this model, the more arbitrageurs that trade in an asset, the closer its price is to the fundamental value and the less attractive it is for each of them to trade in that asset. As a result, arbitrageurs spread themselves out between assets and, in equilibrium, each asset is somewhat mispriced.

In period 1, payoff uncertainty is resolved for short-term assets and prices move to fundamental values. The same happens for long-term assets in period 2. When the price moves to the fundamental value, arbitrageurs settle their accounts and trading ends.

---

[1] In our simple specification, the opportunity cost of tying up one's funds is not modeled. Introducing it strengthens the results.

In equilibrium, returns from arbitrage in each security must be the same. In particular, the number of arbitrageurs investing in each security of a given type is the same, provided that S_i hitting each security of that type is the same. Denote the respective numbers of arbitrageurs by n_s and n_l.

Consider the return from investing in a short-term asset. In period 0, the arbitrageur buys $b/[V_s - S_s + n_s b]$ shares of that asset with borrowed b dollars, receives in expectation V_s for each share in period 1, and must in period 1 repay bR to his lenders. As of period 1, the value of this investment is equal to

$$b[V_s/[V_s - S_s + n_s b] - R]. \tag{4}$$

If the arbitrageur invests in the long-term asset, he buys $b/[V_l - S_l + n_l b]$ shares of that asset in period 0. In period 1, he does nothing and in period 2, he gets the average of V_l per share and repays bR^2. As of period 1, the value of this investment is equal to

$$b[V_l/(R[V_l - S_l + n_l b]) - R]. \tag{5}$$

The difference between (4) and (5) is that, in (5), capital gains are deferred and discounted at the rate R since the arbitrageur is at a borrowing constraint. As R rises, arbitrage in the long-term asset becomes less lucrative and eventually unattractive.

In equilibrium, the returns to arbitrage on the long- and the short-term assets must be equal. Setting (4) equal to (5) we get

$$\frac{V_l}{V_l - S_l + n_l b} \cdot \frac{1}{R} = \frac{V_s}{V_s - S_s + n_s b}. \tag{6}$$

The fraction on each side of (6) is the ratio of value to price of long- and short-term assets, respectively. Since $R > 1$, (6) says that, in equilibrium, the long-term asset is more underpriced in percentage terms (when the noise shock is negative) than the short-term asset. To induce the marginal arbitrageur to invest in the long-term asset, it must be more underpriced than the short-term asset. This result is independent of the values of noise shocks to the two types of assets, or of the relative attractiveness of long- and short-term projects. If we raise total arbitrage resources (raise the total number of arbitrageurs or b), more of them flow into each asset and the mispricing of each asset is reduced, but the greater relative mispricing of long-term assets continues.[2]

III. The Supply of Assets

The compensation of managers of firms with traded equity typically depends in part on short-term equity performance. Poor equity performance raises the likelihood of replacement, either by the board or through a hostile takeover (see our 1989 paper with Randall Morck). By comparison, increases in managers' incomes from good stock performance are probably small. In this respect, temporary overpricing of the stock is much less of a benefit to managers than temporary underpricing is a cost.

This reasoning suggests that managers will choose short over long-term investment projects, since picking the latter allows their equity to be more mispriced in equilibrium, *ceteris paribus*, and threatens their jobs. If the probability of replacement varies directly with the percentage underpricing, then all managers who are not also large, long-term shareholders would choose short-term investment projects on the margin. In this way, short terming by rational arbitrageurs leads to short terming by the firms. If $V_l > V_s$, managerial short terming is socially inefficient, since the economically more valuable long-term projects are priced less accurately in the stock market and therefore avoided.

If a manager does undertake a long-term project and his firm becomes underpriced, a hostile acquirer can profit by changing the firm's investment policy. A hostile acquirer *must* make a change to profit from his acquisition fast, for otherwise an acquisition is the same as long-term arbitrage and there is no reason to buy the whole firm at a premium. The changes that a hostile acquirer can make are of two sorts. He can remove the reason that some investors do not like the firm. For example, if the stock market does not like conglomerates, he can bust up the firm and sell off the pieces. More in the spirit of our model, the hostile acquirer can shift the firm's cash flow toward the present by cutting investment, raising dividends, and selling some of the divisions. Because these changes reduce uncertainty and the likely duration of mispricing, they attract more arbitrageurs and reduce underpricing, even if the underlying noise trader pessimism about the firm remains. After the acquirer makes these changes, he can sell the firm to the public at a higher price since underpricing caused by noise is now less significant.

---

[2] Similar results obtain if arbitrageurs are risk averse and long-term arbitrage is more risky than short-term arbitrage.

Of course, this is precisely what some hostile acquirers as well as leveraged buyout specialists do. They first effectively reduce the duration of the firm's assets by cutting investment and selling off divisions, and then sell the firm to the public. Michael Jensen (1986) enthusiastically interprets such evidence as pointing to elimination of inefficient investment pursued by corporate managers dedicated to growth. Our model suggests, in contrast, that some of the investment cuts eliminate good long-term projects that have previously been responsible for greater stock market underpricing.

In this way, takeovers substitute for long-term arbitrage. With long-term arbitrage, the smart investor must wait until the elimination of mispricing to realize the returns on his trade. When a firm pursues a long-term investment project, the wait can be long. With a hostile takeover, in contrast, the smart investor can gain control and turn a long-term asset into a shorter-term asset that is less mispriced by the market. It is sometimes said that smart money, unless they are willing to wait a long time to profit from an investment, must be ready to take over the firm. Our model explains how this is so.

IV. Self-Fulfilling Short-Term Arbitrage

We have assumed throughout that participation of each arbitrageur brings the period-0 prices closer to their fundamental values, but has no effect on whether prices return to fundamentals in period 1 or period 2. As a result, participation by each arbitrageur reduces the returns from arbitrage to others. However, participation by each arbitrageur can also *raise* the attractiveness of arbitrage to others by bringing prices to fundamental values *faster*, and so reducing the time it takes others to profit from their own detection of mispricing. If the price returns to fundamentals not through a public news event but through the work of arbitrage, then each arbitrageur will trade more aggressively on a piece of information when he knows that other arbitrageurs will later detect the same mispricing and eliminate it more rapidly. Conversely, it does not pay to trade in an asset which no one else follows and which can therefore remain mispriced forever.

In addition, participation by more arbitrageurs reduces noise trader risk, since each smart trader knows that if noise traders become even more biased in the future, more smart traders will lean against them. Since price fluctuations in the asset are smaller when more traders participate, arbitrage is less risky for each of them. To be sure, it does not pay to be in a market with so many arbitrageurs that no mispricing occurs, but at the same time countering the mispricing of a security by oneself might be too risky.

These arguments provide a further reason why arbitrage will focus on short-term assets. When a lot of traders are arbitraging away price discrepancies in short-term assets, arbitrage in them becomes both faster and less risky. This presumably makes the attractiveness of arbitrage to each trader even greater. In contrast, because few traders arbitrage the mispricing in long-term assets, such mispricing might take even longer to correct as the forces bringing prices to fundamentals are weak. Moreover, the arbitrage in the long-term assets will be very risky, since so few traders will be countering future waves of noise trader sentiment. With positive externalities between arbitrageurs, the greater mispricing of long-term assets relative to short-term assets can be self-fulfilling. Almost all arbitrageurs will focus on the same assets and events, such as takeover or earnings announcements, and few will commit time and resources to long-term arbitrage.

V. Conclusion

Arbitrage generally serves the useful social function of bringing asset prices closer to fundamental values. Arbitrage itself, however, is guided by maximization of arbitrage profits. We have shown that the private costs and benefits of arbitrage lead to its clustering on the trading of short-term assets. This clustering in turn leads to systematically more accurate pricing of short-term assets than of long-term assets, even though efficient capital allocation and managerial evaluation might be better served by the opposite bias.

REFERENCES

De Long, J. Bradford et al., "Noise Trader Risk in Financial Markets," *Journal of Political Economy*, forthcoming 1990.

Jensen, Michael C., "The Agency Cost of Free Cash Flow, Corporate Finance and Takeovers," *American Economic Review Proceedings*, May 1986, *76*, 323–29.

Morck, Randall, Shleifer, Andrei and Vishny, Robert W., "Alternative Mechanisms for Corporate Control," *American Economic Review*, September 1989, *79*, 842–52.

Stein, Jeremy, "Takeover Threats and Managerial Myopia," *Journal of Political Economy*, February 1988, *96*, 61–80.

______, "Efficient Stock Markets, Inefficient Firms: A Model of Myopic Corporate Behavior," *Quarterly Journal of Economics*, November 1989, *104*, 655–70.

Stiglitz, Joseph E. and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

The Efficiency of Equity in Organizational Decision Processes

By PAUL MILGROM AND JOHN ROBERTS*

It is now widely accepted that *rent-seeking* (Anne Krueger, 1974) or *directly unproductive profit seeking* (Jagdish Bhagwati, 1982) may cause inefficiencies in the context of public sector decisions. The possibility that government decisions (for example, about taxes, quotas, franchises, or standards) may create or redistribute rents induces private parties to spend valuable resources to influence that distribution, even when such expenditures carry no social benefit. Moreover, the social cost of rent seeking can exceed the value of the resources spent trying to gain and protect rents, for example, because the fear of losing wealth through redistribution reduces the incentives for wealth creation.

Treating rent-seeking activities as characteristic of public sector decision processes ignores the fact that similar phenomena are to be found in firms, unions, and other private sector organizations. Our recent work (Milgrom, 1988; our 1988, 1990 papers) attempts to identify the advantages of decision processes (private or public) that permit rent seeking, and to incorporate these into a cost-benefit analysis of optimal decision processes.

Our work begins with an analysis of why rents and quasi rents arise in organizations and of the forms that the rent seeking they engender may take. Some measures to insulate the decision process from rent seeking, such as limits on the provision of information by interested parties or restrictions on the range of options considered, may degrade the quality of decisions, especially by blocking the flow of valuable information. Optimal decision processes balance the costs of rent seeking against the value of information obtained.

Several aspects of the rules affect the opportunities that members have to spend resources trying to alter the distribution of rents. To ascertain the possibilities for rent seeking, the analyst must ask questions like: Can the parties propose new initiatives at any time? Can they give volumes of testimony in a form of their own choosing? Can they appeal adverse decisions? Are decision makers obliged to respond to the parties' initiatives? Is the range of actions that they can take in response relatively broad, rather than being tightly constrained by property rights or other formal rules? More affirmative answers to these questions lead to more opportunities or greater incentives for costly rent seeking. As a kind of shorthand, we call decision processes that have more of these elements "more open" processes.

The very elements that make a process open to rent seeking may also add flexibility and responsiveness, helping to ensure that important ideas and proposals are fully considered. From this perspective, the benefits of openness can (in principle) be measured by a "value of information" calculation, in the usual manner of statistical decision theory. Weighing the costs and benefits, it follows that a more open process is more desirable when the rents available for redistribution are low, and the value of the information that might be acquired is high. Conversely, when the potential for redistribution is high and the value of information is low, the optimal decision process is less open.

This kind of reasoning helps to illuminate the variations in the decision processes that are found in many organizations. We have discussed a number of examples in our earlier work, including the characteristics of personnel departments, the contrasting patterns of decision making, employment and compensation in U.S. and Japanese firms, and the frequently manifested policy of limiting observable pay differentials.

---

*Professor of Economics, Stanford University, and Jonathan B. Lovelace Professor of Economics, Graduate School of Business, Stanford University, Stanford, CA 94305, respectively. This work was supported by the National Science Foundation.

As a further illustration, consider the way tenure decisions are made at Stanford University. The tenure appointment decision is one of the most important ones faced by the University, since mistakes in tenure appointments tend to remain in place for many years. For the party whose tenure is being decided, and for any untenured faculty member whose potential slot may be filled, the redistributive consequences of this decision may be very large. For the tenured faculty and those from other departments, the redistributive potential is much smaller. Accordingly, the detailed discussions of tenure decisions are mostly limited to other tenured faculty and are reviewed by a University Committee with representatives from other departments. Because of its importance, a successful tenure case will typically incorporate evidence of various kinds, the main restrictions being that sufficient information is supplied to make an accurate judgment and that outside experts in the candidate's field be consulted. For those who participate in this decision, the form of the permitted inputs are not highly structured; reports and letters can raise almost any issue that the writer deems relevant. But, for those most personally affected by the decision, an attempt is made to conceal the identities of the decision makers and to limit access to the decision process.

According to our theory, it is desirable to employ different sorts of processes for decisions with large or small redistributive consequences. It would then seem to follow that organizations will try to separate efficiency and redistributive issues, and to settle the latter issues first, so that they won't interfere with decisions about how best to create value. To some extent, this theoretical prediction is borne out in reality. At Stanford, for example, student housing priorities are fixed by lottery and then the students are permitted to choose from the housing that remains after higher priority students' choices have been filled. Similarly, airline flight attendants are given priority to bid for routes according to seniority. In both cases, the distributional issue is treated mechanically (by lottery or seniority rules), and the remaining efficiency aspect is delegated to the parties with the relevant information.

Nevertheless, in many situations, the expedient of separating the distributional and efficiency issues is not available or efficacious. For example, although some airlines employ the same process for assigning flight crews to routes as they do for cabin crews, this can be costly, as when the least-experienced pilot, copilot, and flight engineer all end up together on the same run that no one with more seniority wanted. A wage policy that provides compensating differentials to losers from a decision might also sometimes immunize value creation from distributional considerations. For several reasons, however, the policy of compensating differentials is not generally feasible. First, as we have argued elsewhere, the compensation system may be serving other important objectives, such as spreading risk, retaining key employees, motivating hard work, encouraging investment in human capital, and so on. Second, it is not generally possible to know how much compensation would make an individual neutral about a decision. Finally, the use of cash payments opens new opportunities for extortion, bribery, and other forms of costly rent seeking. For all these reasons, organizations must often manage decisions that have some significant distributional effects. Such situations are our concern here.

The theoretical model sketched below considers the process of making proposals for new activities or investments in an organization where a separation between efficiency and distributional issues cannot be achieved, and where the main providers of proposals, information, and analysis are the very same parties who are affected by the decisions. Under these circumstances, we show it is often optimal for the decision maker to commit to a policy of assessing alternatives according to their distributional impact or equity as well as to their contribution to the organization's overall objectives. The reason for this is that requiring some "fairness" in the distribution mitigates the incentives for rent seeking, and also may lead to better overall proposals being generated. However, the decision maker must not pursue this concern with equity too far, because to do so risks destroying the incentives for those who propose projects to do their jobs well.

I. A Model of Project Proposal and Evaluation

We model an organization whose objective is to maximize the expected total (transferable) utility of its members. It is, at a minimum, controversial to assign no weight to equity concerns in discussing the organization's true objectives (Margaret Meyer and Dilip Mookherjee, 1987), but we make this assumption with an analytical point in mind. With no weight accorded to "equity" in the organization's exogenous objective, any concern for equity that emerges from the model is attributable to an endogenous concern for efficiency in the processes of generating alternatives and information and of making decisions.

We shall assume that there is a central coordinator for the organization who has the power to make decisions about project adoption, and who is chosen and compensated so that she has no personal interest in the outcome, but rather seeks the organizational goal. However, the members have better information than the coordinator about the possible projects, so it is they who generate proposals for projects that the organization might conduct.

Our main conclusions will be derived using variations of the following simple model. The organization has just two members, only one of whom generates proposals. A proposal is characterized by the utility pair (u, v) for the two members, where u is the utility accruing to the proposer, $(\bar{u}, \bar{v})$ is the default utility if no project is proposed or none is accepted, and 0 is the lower bound on each of the utilities across all possible projects. The proposer can generate projects $(u, v) \geq (0,0)$ at a personal cost $C(u, v)$ that is increasing in each argument. The central coordinator can observe any project (u, v) that is proposed, and has the power to accept or reject it. Assuming differentiability and convexity of C, the first-best is achieved when the proposer selects and the coordinator accepts a project that maximizes $(u+v)-C(u, v)$.

If the coordinator accepts all project proposals that increase the organization's performance $(u+v>\bar{u}+\bar{v})$, then the proposer will find it most profitable to generate proposals that maximize u subject to the acceptance constraint, rather than to generate proposals that maximize the organization's objective $u+v$. It is possible that this leads to no loss for the organization (if $C(u, v)$ takes the form $C(u+v)$) but, in general, u will be set too high and v too low compared to the organization's optimal levels (u^*, v^*). We shall suppose that to discourage this sort of choice, the coordinator has the power to commit herself to an acceptance policy, under which, for example, she accepts only proposals with $v \geq v^*$. In this simple model, the acceptance policy just described would lead to the first-best outcome for the organization provided that $u^*-C(u^*, v^*) \geq \bar{u}$, so that the proposer is willing to make a proposal.

A somewhat richer description of the optimal policy is obtained when we vary the model as follows. Suppose that the proposer cannot pick a project (u, v) exactly, but instead that when he aims for (u, v) and incurs costs $C(u, v)$, he gets a project $(\hat{u}, \hat{v})$, where the conditional distributions of $\hat{u}$ and $\hat{v}$ are $f_u(\hat{u}|u)$ and $f_v(\hat{v}|v)$, and where f_v exhibits the monotone likelihood ratio property, so that higher realizations of $\hat{v}$ are evidence of higher values of v. Suppose further that the proposer can conceal any project proposal, so that the only proposals that will be forthcoming are those with $\hat{u} \geq \bar{u}$. Finally, suppose that the coordinator can commit to an *acceptance policy* X according to which any proposal $(\hat{u}, \hat{v})$ that is made is accepted with probability $X(\hat{u}, \hat{v})$. With these assumptions, we reach the following basic conclusion.

PROPOSITION: (*Apparent Concern for "Equity"*) *The optimal project acceptance policy will accept or reject each project without randomization* ($X=0$ *or* 1). *The only acceptable proposals are those with a high enough level of utility* $\hat{v}$ *for the nonproposer, where the* *requisite level may depend on* $\hat{u}$. *(That is, if* $X(\hat{u}, \hat{v}) = 1$ *and* $\hat{v}' > \hat{v}$, *then* $X(\hat{u}, \hat{v}') = 1$.)[1]

This first model captures one of the efficiency-enhancing roles of equity concerns, but does not highlight the expenditure of resources in redistributive efforts that is typically labelled "rent seeking." To focus on that aspect of the decision problem, we return to the case where the proposer can choose (u, v) exactly, but we add the hypothesis that success in generating any proposal at all is not assured. The probability of generating a successful proposal may depend, in general, on the proposer's efforts p in searching out and developing the proposal and the other member's efforts q in blocking it.[2] More specifically, we assume:

(A1) The efforts of the two members to generate and block proposals are unobserved, but the actual proposal (u, v) is observed by the coordinator if it is made.

(A2) The probability that a successful proposal is made is $P(p, q)$, where p is the search effort of the proposer and q is the blocking effort of the other member. We assume that P is increasing in p and decreasing in q.

(A3) The cost functions $C(u, v, p)$ of the proposer and $K(q)$ of the other member are both smooth, increasing, and convex.

(A4) The proposer seeks to maximize his own personal utility, which is $P(p, q)u + [1 - P(p, q)]\bar{u} - C(u, v, p)$, and the other member seeks to maximize his, which is $P(p, q)v + [1 - P(p, q)]\bar{v} - K(q)$.

The central case for analysis is one in which the division of benefits from the proposal do not influence costs (i.e., $C(u, v, p) = C(u + v, p)$), so that incentive considerations alone determine how the benefits of the project ought to be shared between the two members. We call this the *cost-neutral* case. Let (u^*, v^*, p^*, q^*) be a "first-best" solution of the proposal problem, that is, a combination that maximizes $P(p, q)(u + v) + [1 - P(p, q)](\bar{u} + \bar{v}) - C(u, v, p) - K(q)$ without regard to the incentive constraints.

PROPOSITION: (*Acceptance Policy in the Cost-Neutral Case*) *Suppose that the proposer gains in moving from the status quo to the first-best* $(P(p^*, q^*)(u^* - \bar{u}) - C(u^*, v^*, p^*) \geq 0)$. *Then, in the cost-neutral case, the optimal acceptance policy achieves the first-best. The optimal policy consists of accepting just those proposals which set* $v \geq \bar{v}$.

In this cost-neutral case, the optimal acceptance policy has several effects. First, allowing the proposer to enjoy all the net benefits of the proposal aligns his interests with those of the organization and induces him to work hard at generating proposals. Indeed, with v fixed at $\bar{v}$, the proposer's choices of p and u will coincide exactly with the organization's most preferred choices, regardless of the cost function C. This makes it possible to delegate broad authority to the proposer subject only to the "equity constraint" that new proposals must not take anything away from other members. A second effect of this acceptance policy is that the other member is assured he will not be damaged by the new proposal, so that he will not resist it ($q = 0$). The need to avoid excessive resistance to change is an important aspect of any real organizational decision system.

Using this central case as a basis, it is easy to examine related models in which some one aspect of the formulation is varied. For example, if the active cooperation of the other member can help to increase the probability P of a successful proposal, then the optimal v is larger than $\bar{v}$. In a related vein, H. Lorne Carmichael (1988) has explained the institution of tenure as an attempt to induce the active cooperation of existing faculty in selecting other outstanding new faculty. He argues that without the guarantees of the tenure system, existing faculty might feel their own positions threatened, and so might prefer to hire mediocre colleagues. A second variation replaces the assumption that C depends only on p and $u+v$ with the related assumption that C is symmetric in (u, v) and *strictly* convex; the upshot is that it pays to set $u>v>\bar{v}$ in order to control costs. When the "other member" is an untenured junior faculty member, a different argument applies. Thus, suppose that C_v/C_u and K'' are large, so that it is costly to raise other member's utility v without damaging the senior faculty and the untenured member has little ability to interfere. Then it will be optimal to set v lower than $\bar{v}$, that is, to set a high standard for tenure.

---

[1] This result can be proved as follows. First, solve the "relaxed optimization problem" in which it is assumed that the proposer needs only to be provided with incentives to set v high enough. Then, show that the solution to the relaxed problem has the hypothesized form. Finally, verify that the solution so obtained satisfies the other incentive constraints as well, and is therefore a solution to the original problem.

[2] The role of employees actively blocking the employer's interests is also discussed by Edward Lazear (1989), who investigates how pay equity can alleviate the problem.

An especially important variation on our model arises where the other member expends the organization's time and resources, rather than his own, in attempts to block an unfavorable decision. In terms of our formulation of the problem, this implies the $K(q)$ term remains in the organization's objective function but disappears from the other member's function. It is evident that protecting the other member becomes even more important in that case, for two reasons. First, as in the earlier model, the objections of the other member do double damage: they not only block the organization's decisions but they also add to its costs or detract from its performance. Second, because the cost of blocking is not borne personally by the other member, he is much more willing to set q high, that is, to oppose vigorously any proposed project that might damage his welfare. So, the extent of the damage done and costs borne may be greatly increased in this case.

In practice, of course, employees as members do use the organization's resources to fight personal battles: they take the organization's time to write memoranda, have them word processed by secretaries, and have them read and responded to by higher level managers. To the extent that the personal costs of these activities are much lower than the organizational costs, a failure to protect established interests with an "equity" provision in the acceptance policy can be very costly.

II. Extensions and Further Discussion

An interesting and important extension of the basic model is to the case where there are two or more proposers among the membership of the organization.[3] Some new features emerge in that analysis, and we summarize them here.

In terms of the model just studied, the main change is that the utility associated with a failed proposal is no longer specified exogenously as some $(\bar{u}, \bar{v})$, but is instead endogenous, being determined as a function of the proposals and efforts of others. In a cost-neutral model with competing proposers, it is still true that a first-best outcome can be achieved by setting the other member's utility sufficiently high, but in this case the correct level is the expected utility of the other member if the proposal were to fail. With that standard set for each proposer, one can again show that each proposer's objective corresponds exactly to the organization's objective, and that no other member has an incentive to expend personal resources to block or defeat another's proposal.

Here our formal models deal only with the proposing of projects, but the same general logic applies to their evaluation and implementation. Interested parties will tend to develop positive information and enhancements to support the projects that serve their own interests, neglecting projects that serve the larger interests of the organization. By favoring "equitable" projects, the coordinator can remove some of the bias that naturally afflicts project evaluations. A related but different point can be made concerning the implementation of projects and changes. If, as is often the case, the cooperation of many affected parties is important to make a new project a success, then there is a danger of some of the parties resisting the current project in the hopes of implementing an alternative that is more in accord with their interests. A policy that emphasizes equity tends to limit this kind of resistance. Also, by allowing those responsible for successful implementation to get a larger share of the benefits, the coordinator can encourage active cooperation with the proposed projects.

---

[3] Our 1986 paper showed in a somewhat different context how competition between proposers could be used to overcome severe informational disadvantages.

REFERENCES

Bhagwati, Jagdish, "Directly Unproductive, Profit-Seeking (DUP) Activities," *Journal of Political Economy*, October 1982, *90*, 988–1002.

Carmichael, H. Lorne, "Incentives in Academics: Why is There Tenure?," *Journal of Political Economy*, June 1988, *96*, 453–72.

Krueger, Anne, "The Political Economy of the Rent-Seeking Society," *American Economic Review*, June 1974, *64*, 291–303.

Lazear, Edward, "Pay Equality and Industrial Politics," *Journal of Political Economy*, June 1989, *97*, 561–80.

Meyer, Margaret and Mookherjee, Dilip, "Incentives, Compensation, and Social Welfare," *Review of Economic Studies*, April 1987, *54*, 209–26.

Milgrom, Paul, "Employment Contracts, Influence Activities and Efficient Organization Design," *Journal of Political Economy*, February 1988, *96*, 42–60.

______ **and Roberts, John,** "An Economic Approach to Influence Activities and Organizational Responses," *American Journal of Sociology*, July 1988, Suppl., *94*, S154–79.

______ **and** ______, "Bargaining Costs, Influence Costs and the Organization of Economic Activity," in J. Alt and K. Shepsle, eds., *Perspectives on Positive Political Economy*, Cambridge: Cambridge University Press, 1990.

______ **and** ______, "Relying on the Information of Interested Parties," *Rand Journal of Economics*, Spring 1986, *17*, 18–32.

Asymmetric Information and the New Theory of the Firm: Financial Constraints and Risk Behavior

By BRUCE C. GREENWALD AND JOSEPH E. STIGLITZ*

Recent developments in the theory of the firm have been closely related to developments in examining the economic implications of imperfect information. This treatment of the firm goes beyond both its Arrow-Debreu incarnation as a disembodied production set, and its formulation as an efficient nexus for minimizing the cost of economic transactions due to Coase, and more recently to Oliver Williamson (1985). Asymmetrically distributed information between a firm, as employer, and its workers has replaced the traditional view of a firm that hires labor at fixed (or monopsonistically increasing) wages in well-defined labor markets with one in which firms actively manage long-term employment relationships, on average pay wages in excess of those available in the labor market at large, control workers with carefully designed incentive mechanisms, and often ration access to jobs.

Similar asymmetries of information between outside investors, who provide capital, and inside managers, who control its use, have led to comparable developments in the theory of how firms acquire and deploy capital. In both cases, there have been two dimensions to the theoretical developments involved. On the one hand, much recent attention has been paid to the internal structure of the firm—how rewards for individual workers should be designed, what constitutes appropriate hierarchical or reporting structures, and how the internal quality of life of the firm depends on these factors. On the other hand, a new view of the external aspect of the firm has developed—how it is likely to react to external environmental and policy changes. The focus of this paper is on the latter external dimension since developments in this area have begun to lead to a broadly common set of implications for firm behavior.

Among these developments, those with perhaps the most striking implications for external firm behavior have arisen from a reexamination of the role of financial variables (leverage, cash reserves, financial strength, etc.). The classical Modigliani-Miller approach to financial policy concluded that the financial structure of a firm was irrelevant to both its value and its operating decisions, and the neoclassical theory of the firm assumed further that its financial position was irrelevant. Yet both informal observations and systematic empirical evidence have suggested that financial structure and position are of considerable importance to firm behavior. For example, notions of deep pockets or financial strength, which have always played a significant role in qualitative observations of firm behavior, have no place in a classical Modigliani-Miller world. Nevertheless, recent empirical studies of investment have demonstrated the importance of financial variables.

Models of imperfect information in financial markets have altered the traditional view in two important ways. First, if information is asymmetrically distributed between the buyers and sellers of financial instruments, then certain financial markets, such as that for equities, may break down or be severely limited (a form of the lemons problem identified by George Akerlof, 1970), and accordingly the free access to all forms of financing envisaged by Modigliani-Miller may not exist. In loan markets, there may be credit rationing. In these cases, financial structure and position matter and affect firm behavior. Second, if information is asymmetrically distributed between those who make decisions (agents) and the theoretical beneficiaries of those decisions (principals), then the reward functions that govern firm decision making may not have the form of simple valuation maximization assumed in the neoclassical theory. This paper seeks to examine the consequences of both kinds of imperfections for the behavior of firms and their evolution over time. In doing so, it looks at an exemplary case of imperfect competition and explicit investment in long-run research and development. However, the lessons of that model point more broadly to the general direction in which informational imperfections in financial markets affect the external dimension of firm behavior.

---

*Bell Communications Research, Morristown, NJ 07960, and Stanford University, Stanford, CA 94305, respectively.

I. The Model

We will assume for expository purpose that firm decision making takes place in two distinct stages. In the first, the firm decides on a level of productivity improving activity, x. This may be interpreted as either explicit *R&D* spending or as spending on an overhead establishment whose task is to improve the firm's productivity, broadly construed to include the efficiency of its marketing and general administrative activities. The activity reduces the constant marginal cost of output with which a firm enters the second stage of market activity, in which output is produced and sold. This constant marginal cost c will be assumed to depend on an inherited level of production knowledge, c_o, that depends on the spillover benefits of past economywide productivity improving activities; on first-stage overhead expenditures to improve productivity, x, which we will refer to for convenience as *R&D*; and on a random factor s, with expected value one, which determines the success of those efforts: $c = sg(x)c_o, E(s)=1$, where the function g is assumed to be decreasing in x.

In the second stage of market activity, firms maximize profits, given their cost levels and the demand for their product, d, which we assume depends on both the firm's own price p, and the average price $\bar{p}$ charged by other firms. Formally, therefore, in the second stage, firms $\max \pi = (p-c)d(p,\bar{p})$, which leads to an optimal price and profit level: $p^* = mc, m = 1/(1-1/e)$ and $\pi^* = (m-1)cd(mc, m\bar{c})$, where $\bar{c}$ is the average of the cost of other firms and e is the own-price elasticity of the demand function, d, assumed to be a constant.

The payoff function to first-stage *R&D* activity is, therefore,

$$\pi(x) = (m-1)\,sg(x)\,c_o d(msg(x)\,c_o, m\bar{c}).$$

If there are a large number of competing firms, then $\bar{c}$ will be approximately equal to $\bar{g}c_o$, where $\bar{g}$ is the average of $g(x)$ over other firms and c_o, the initial inherited technology, is assumed to be the same for all firms (i.e., technologies fully spill over into the public domain after a single period). The second-stage profit function can, thus, be written

$$\pi(x) = (m-1)\,sg(x)c_o d(sg(x),\bar{g}),$$

where the m and c_o terms in the demand function have been suppressed for notational convenience. Profits are a random variable looking forward from the beginning of the first stage in which x has been chosen.

The choice of a level of research activity in stage one depends on the objective function of a firm at that time and, in particular, on its attitude toward risk. We assume that the firm in question is owned and managed by an individual who maximizes the utility of end-of-period wealth, having sold a fraction $(1-a)$ of the firm to outside investors. Thus, in stage one, the firm decision maker maximizes expected utility:

$$(1) \qquad \max_{x} E\left[u\left(w+(\pi(x)-x)a\right)\right],$$

where we assume that u is characterized by decreasing absolute risk aversion and w, initial wealth, includes the proceeds from the sale of the fraction $(1-a)$ of the firm. For the case in which the risks associated with x are unsystematic, such an entrepreneur, in a world of perfect information, would diversify completely and simply maximize the expected value of $\pi(x)$. However, if he both is and is known to be better informed about his own prospects for productivity improvement than investors at large, then the firm's owner/manager will be constrained to hold an excessive fraction of the firm in order to signal confidence in his own prospects. (See Hayne Leland and David Pyle, 1977.) And firm behavior, even with respect to the unsystematic risks that are associated with active productivity improvements, will be governed by the risk aversion implied by the utility function in equation (1).

An objective function similar to that of equation (1), or at least with similar implications, emerges from a wide range of imperfect information models. For example, the distinction between owner-managed and professionally managed firms is not especially significant here. When professional managers' actions are unobservable, in effect managers become part owners of their firms' equity capital. The managers' wealth then consists of their private holdings, w, plus the fraction a of the terminal equity (profit) of the firm that they are able to appropriate to their own use. Since this latter part of their wealth is nonfungible, and since the managers, like owners, should be deterred from issuing shares by informational considerations, managers will be maximizing a function similar to that of equation (1). Alternatively, if agency arrangements are made with risk-averse managers and these arrangements are limited for practical reasons to payoffs that are linear in the profits of the firm, an objective function identical to that of equation (1) will emerge, with the variable a now representing the slope of the payoff function and w now being the returns to safe projects (i.e., liquid investments).

Similarly, since the fact that issuing shares has a negative impact on the market value of the firm makes the firm reluctant to issue shares, if external funds are raised, it will in theory be (and in practice is) predominantly through loans. But this makes the firm face a risk of bankruptcy, a risk that is affected by firm behavior. We can show that the behavior generated by maximizing expected profits minus an expected cost of bankruptcy (the cost of bankruptcy times the probability of bankruptcy) is similar to that generated by equation (1) under plausible restrictions on the firm's cost and probability of bankruptcy. Thus, if professional managers are subject to an agency arrangement that either explicitly or informally rewards them with a share of profits, but imposes a large penalty in the event of bankruptcy (i.e., dismissal with a stigma that significantly impairs future earnings), then the resulting objective function produces behavior almost exactly identical with that of equation (1),[1] when (as we assume here) the utility function is characterized by decreasing absolute risk aversion.

The first-order condition for the problem of equation (1) can be written

$$c_o\bar{d}(-g_x) = \bar{d} + \mathrm{cov}(u', ds) + \mathrm{cov}(ds), \tag{2}$$

where E has been normalized to equal one, both covariances are negative (since better outcomes for research activities are correlated with higher sales and sale revenues), and $(-g_x)$ is positive. The right-hand side of this equation is the risk-adjusted cost of research investment at the margin. It is least when the firm is risk neutral (i.e., $\mathrm{cov}(u'd, s) = 0$), and increases with firm risk aversion. The left-hand side is the marginal return to research expenditures. Assuming the appropriate second-order condition is satisfied, the behavior of a finance and/or agency constrained firm can be derived straightforwardly from equation (2).[2] Broadly speaking, these firms act as if they are risk-averse individuals, which is not surprising since one basic impact of imperfect information is to interfere with the proper distribution of risk.

A first consequence of this is that an increase in the wealth of firm owners (financial strength of the firm) leads to reductions in the size of the covariances and hence a reduction in the risk premium associated with overhead investment. Thus, as a rule, increases in historical firm profitability and/or temporarily high current profitability (increasing w) lead to higher levels of research investment (or investment generally), more rapid expected productivity growth for each firm, and, through future spillover effects, higher productivity for the economy as a whole. Consequently, temporary demand disturbances which reduce current profitability will lead to effective supply contractions (as in a standard Keynesian model) as firms reduce investment activity (*R&D* etc.) and these, in turn, lead to permanently lower future output. Not only is the informal notion of the importance of the strength of a firm's financial position validated in these imperfect information models, but they also provide a mechanism for the propagation of macroeconomic demand disturbances and the persistence of their effects.

---

[1] See our 1987 paper. This requires that bankruptcy costs increase with the size of the firm (as they appear to do in practice).

[2] The discussion below assumes implicitly that the ownership factor a does not change with the exogenous variables. In the pure finance-constrained interpretation, this is appropriate. In the agency interpretations, a may change. However, the basic comparative static properties of the model are unaltered.

Increases in the uncertainty of the technological or future market environment (i.e., in the variance of s) have a similar impact. The risk premium associated with investment rises, even when the risks involved are unsystematic, and the level of research investment falls. Ultimately, therefore, productivity growth declines across the economy as a whole. The imperfect performance of financial mechanisms for risk distribution means that an economy may react negatively to increases in variables like relative price uncertainty (that affect the uncertainty of returns to productivity improvement). These are all, it should be noted, effects that are not part of the responses of the traditional firm.

II. Dynamics

Another significant aspect of the finance constrained firm is that it is characterized by a well-defined dynamic evolution, that depends on the evolution of the owner/manager's wealth (or in the case of a managerial firm, the accumulated retained earnings of the firm). The wealth of a firm's owner in period t is determined by the relationship

$$(3)\quad w_{t+1} = a\pi(x_t) + (w_t - ax_t)(1+r_t) - y_t(w_t),$$

where r_t is a safe rate of return on assets not invested in the firm at stage one, y_t is consumption in period t that is an increasing function of wealth in period t, and, as described above, x_t is the level of investment in research and development which depends positively on w_t; namely, $x_t = h(w_t)$, $h' > 0$.[3] For an individual firm, equation (3) is a stochastic difference equation, since $\pi(x_t)$ is a random variable, depending on the random return variable s_t. Under reasonable circumstances (a bounded distribution of s, sufficiently rapidly diminishing returns to x and sufficiently concave preferences) a limiting distribution of w_t, independent of initial wealth, will exist.[4]

Among other factors, this limiting distribution will depend in the present model on the wealth levels of other firm owners. Higher economywide wealth levels imply higher economywide levels of *R&D* spending, lower expected costs and hence prices for competing firms, lower expected demand at each own-price level for any single firm, and therefore, lower expected profits for each firm. Thus, for the economy as a whole, there will be in equilibrium (again under reasonable circumstances) a limiting distribution of wealth across firms which implies a limiting distribution across firms of research and development expenditures.

Finally, the long-run evolution of the model depends on the evolution of c_o that embodies the inherited technology in each period. In order to simplify matters as far as possible, we will assume that the demand function at t, $d_t(p_t, \bar{p}_t)$, has the form $d_t(p_t, \bar{p}_t) = (1/c_o(t))d(p_t/\bar{p}_t)$.

As inherited costs fall, real income rises, demand increases proportionately and the share of demand served by each individual firm depends on its price relative to the average price of all other firms. Under these conditions, $\pi(x_t)$ and the level of x_t chosen by each firm is independent of $c_o(t)$. Thus, the analysis of the dynamics of the model can be separated into 1) an analysis (described above) of the evolution of the wealth and *R&D* expenditures of individual firms, and 2) an analysis of the overall improvement in the baseline technology of the economy, embodied in c_o, as a function of the wealth and *R&D* spending distributions. Formally, that second step can be described by an equation of the form $c_o(t+1) = c_o(t)G(\bar{x}_t)$, which describes the process of technology spillovers where $\bar{x}_t$ is a vector whose elements are the levels of x_t for the individual firms.

---

[3] In a more general model, we can derive the equilibrium $y_t(w_t)$ function also.

[4] This holds as long as the basic firm decision problem is unaltered by changes in c_o; which requires that the demand function facing each firm increases linearly with $1/c_o$ (i.e., real income).

In order to trace the long-term consequences of any environmental or policy change within this context, its consequences for the evolution of both the wealth distribution of firm owners and technology must be examined. For example, an unexpected increase in the wealth of firm owners (due, for example, to an unexpected increase in demand) in a single period will not alter the limiting distribution of firm-owner wealth or *R&D* expenditures. It will not, therefore, alter the long-run growth rate of productivity. However, it will temporarily increase *R&D* expenditures, leading to an increase in the level of productivity. Consequently the new limiting growth path will involve a permanently higher level of output.

An increase in environmental uncertainty (due, for example, to greater policy instability or to greater uncertainty in the results of research efforts) will in the short run lead to a reduction in *R&D* efforts at the initial existing levels of firm-owner wealth. However, the decline in economywide *R&D* expenditures will tend to increase the profitability of such efforts by individual firms both on average and at the margin. The marginal effect involved, will, in the short run, partially offset the original impact of greater uncertainty on risk-averse firm-owners. The increase in average profitability will lead to wealth accumulation by firm-owners and, over time, to a partial restoration in *R&D* expenditures. On balance, the increase in uncertainty is still likely to lead to lower *R&D* expenditures and lower growth rates of productivity. However, the extent of the initial decline will be partially offset by subsequent wealth accumulation. Thus, analysis of both the ultimate size and the temporal pattern of the response of the economy to a change in uncertainty (or to a deliberate policy change, like a tax increase) requires that the long-term effects of firm-owner wealth change be taken into account.

III. Conclusion

Imperfect information affects both the internal organization of firms and its external relations with labor, capital, and product markets. The new theory of the firm is built on these foundations. This theory has important behavioral implications that distinguish if from earlier neoclassical theories. This paper, focusing on informational problems in the capital market, including asymmetries of information between providers of capital and firm managers, has argued that as a result firms will act in a risk-averse manner. Several consequences follow: (*i*) the firm will be concerned with its financial structure, and financial structure affects behavior; (*ii*) changes in financial strength (the firm's net worth) have real consequences; and (*iii*) mean-preserving changes in distributions of prices and sales have real effects. Elsewhere (1989) we have shown that the appropriate way to look at the whole set of firm decisions (including those relating to employment, production, pricing, investment that includes inventory changes, and research) is as a dynamic portfolio problem. The results of that model can be contrasted both with the simple neoclassical model of the firm and with attempts to make that model accommodate more of the facts concerning firm behavior, in particular, those postulating costs of adjustment. In some cases, to attain patterns of behavior consistent with the facts (such as the greater variability in quantities than in prices) requires implausible assumptions concerning adjustment costs. But, even in those cases where the neoclassical models and their extensions yield results that are consistent with observation, they fail to account for the facts that firm output and investment spending (in *R&D* and of other sorts) respond directly to changes in firm wealth and to environmental uncertainty of an unsystematic sort, while our models support much of the empirical literature that relates variables like investment directly to firm financial conditions, and more informally, to conditions of uncertainty (due, for example, to high inflation or unstable government policy) in the economy at large (even when these appear unlikely to lead to increases in systematic risk). Here we have used the theory of the finance-constrained firm to explore the dynamics of firm and economy-wide growth. These models have strong policy implications that, while corresponding to widely held informal views, often differ markedly from those based on the neoclassical theory of the firm and its derivatives.

REFERENCES

Akerlof, George, "The Market for Lemons: Qualitative Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, May 1970, *84*, 288–300.

Greenwald, Bruce and Stiglitz, Joseph E., "Toward a Theory of Rigidity," *American Economic Review Proceedings*, May 1989, *79*, 364–69.

______ **and** ______, "Financial Market Imperfections and Business Cycles," NBER Working Paper No. 2494, 1987.

Leland, Hayne E. and Pyle, David H., "Informational Asymmetries, Financial Structure and Financial Intermediation," *Journal of Finance*, May 1977, *32*, 371–87.

Williamson, Oliver, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

STATE AND LOCAL GOVERNMENT FINANCE†

Inflation and the Real Growth of State and Local Government Expenditures

By ROY BAHL AND JORGE MARTINEZ-VAZQUEZ*

Inflation was perhaps the major problem facing state and local government finances as the 1980s began. Double-digit inflation rates throughout much of the late 1970s had driven up the unit cost of providing government services, and because tax bases had not kept pace, an increase in nominal tax rates and cutbacks in public service levels were alleged to have resulted. This paper gives some structure to the discussion of inflation impact on government budgets by formulating a general economic model that separates automatic from discretionary responses, and identifies the relative price, income, and budget effects of inflation. We also estimate the impact of these three components on the real expenditures of U.S. state and local governments during the past two decades.

I. Conceptual and Measurement Issues

Several strands of the public finance literature have picked up on the importance of inflation as a determinant of the growth in government but none, we argue, have asked the most important question: "What is the mechanism by which inflation affects the real demand for state and local government expenditures?"

How does one measure the inflation to which a government will respond? Different deflators are called for depending on whether the purchaser is viewed as a government unit or as a taxpayer (Peter Heller, 1981). There is also the choice between an implicit price deflator or a fixed-weight deflator. The first overestimates the effects of price changes on government purchases and the second underestimates them. Neither is an appropriate deflator for transfer payments.

Most studies of the determinants of state and local government expenditures have used cross-section data, ignoring the inflation issue (Thomas Borcherding and Robert Deacon, 1972). Research on the determinants of long-run changes in government expenditures takes inflation explicitly into account, but these studies have been straightforward empirical testing. Studies for Canada (Richard Bird, 1979) and the United States (Morris Beck, 1985) show that when the government expenditure-GNP ratio is "appropriately" deflated, a smaller real expansion in the public sector is observed. This provides some evidence of a relative price effect, but because no underlying theoretical structure is presented, and because automatic and discretionary changes are not separated, it is difficult to understand how inflation induces changes in government behavior.

Only a few studies have taken on the more explicit objective of comparing the impact of inflation on the revenue and expenditure sides of state and local government budgets. David Greytak and Bernard Jump (1975) measured the potential expenditure and revenue responses to inflation in the 1970s, but under the assumptions that the level and mix of inputs would remain constant, that tax bases would fully respond, and that no discretionary changes would take place. They found that the purchasing power of state and local government revenues eroded by about 10 percent. Attiat Ott (1983) included discretionary as well as automatic changes, adjusted purchases and revenues by "appropriate" deflators, and concluded that inflation, during the 1969–79 period, appeared to generate "tax dividends."

---

†Discussants: Robert P. Inman, University of Pennsylvania; Michelle J. White, University of Michigan; Howard Chernick, Hunter College.

*Department of Economics and Policy Research Program, College of Business Administration, Georgia State University, Atlanta, GA 30303. We are indebted to Robert Inman for a number of helpful comments on an earlier draft, and to Michael Jordan for his research assistance.

Other studies have analyzed the narrower question of the impact of nominal income growth on state and local government revenues and found an elasticity greater than unity (Daniel Feenberg and Harvey Rosen, 1986). John Ross and Richard Reeder (1979) estimated the relationship between revenues and the implicit price deflator, holding constant the nominal GNP gap, and found that revenues were 6–16 percent higher than they would have been in a noninflation world during the 1973–76 period.

Where this research has found, and left, the state of what we know about inflation impacts can be summarized as follows: (a) inflation may induce increases in the relative price of government goods and services, but since government revenues will also be driven up, it is not clear whether the revenue or expenditure stimulation will dominate; and (b) there is some uncertainty about whether (and how) discretionary tax rate changes, and borrowing and expenditure retrenchment should be counted as an impact of inflation.

II. The Model

A stylized model of the local and state subsector is developed here to explore the different ways in which inflation may impact real expenditure growth. We assume two types of agents, voters and bureaucrats. Voters are assumed to be sovereign in that they determine the level of public goods, and, with some time lags, the composition of taxes. Bureaucrats are assumed to minimize the production costs of public goods, and they administer the jurisdiction's debt that is issued to purchase capital goods. Bureaucrats can only deviate temporarily from voters' wishes by, for example, spending rather than rebating federal government transfers. Over the longer run, for given prices and disposable incomes, voter preferences are the only determinant of the jurisdiction's expenditures. This fairly standard median voter model treats grants as a lump sum addition to personal income, and total disposable income is expressed net of federal income taxes.[1]

Given the desired level of public goods, the tax rates are set by the jurisdiction's managers to generate either a balanced budget or a surplus. Because adjustment in tax rates may take place with a lag, and because of the balanced budget constraint and the prohibition against financing current expenditures with debt, it may be necessary for the jurisdiction to supply a disequilibrium level of public goods until tax rates can be adjusted. Such a disequilibrium situation is less likely to occur when the reserve position of the local jurisdiction is strong.

We assume that in production the elasticity of substitution of capital for labor, capital intensity, and the rate of technological change are all smaller for the public good than for the private good and housing (William Baumol, 1967). The initial equilibrium in public and private goods will not be disturbed if all prices and nominal values (incomes, transfers and debt, etc.) continue to increase period after period at the same rate, if the federal income tax is indexed to inflation, and if the real level of grants is held constant. It is highly improbable, of course, that all rates of change in nominal values will be the same in a generalized inflation environment. More likely, inflation is accompanied by changes in the relative price of factors of production and therefore of public vs. private goods (Beck; Bird). The interest in this paper is in this more general condition, and in particular with three effects: (a) an automatic real income effect that comes about because of the progressivity of the federal income tax and the inflow of federal grants; (b) an automatic relative price effect that is due to differential inflation rates for state and local government sector goods and services vis-à-vis housing and other private goods; and (c) the discretionary changes that the income and price effects call forth. Our goal is to disentangle these three effects.

---

[1] A full specification of the model (which includes public goods, private goods and housing) is available upon request from the authors.

Even if a taxpayer's nominal income keeps up with inflation, his real purchasing power can be decreased by a progressive federal income tax that is not fully indexed. Since state and local public services are not inferior goods, this will cause a reduction in the real demand for government revenues. The other component of the real income effect is federal grants, which can partially offset the effects of income tax bracket creep. This model assumes that such offsets occur on a dollar-for-dollar basis and there are no "flypaper" effects associated with grants.

In a generalized inflation environment, it is almost certain that some prices will increase faster than others. In particular, inflationary expectations can lead to wage increases in excess of the general inflation rate. Relatively larger labor shares, lower rates of labor productivity improvement, and lower elasticities of substitution of capital for labor make it much harder for the state and local sector to absorb labor cost inflation. We assume in this model that inflation drives up wages by more than the general rate of price increase, hence the marginal cost of providing government goods increases by more than that for private goods. This increase in the relative price will lead to a decrease in the demand for government goods and services. Whether this will be accompanied by an increase in desired expenditures depends on the price elasticity of demand.

The third effect is the impact of discretionary actions and institutional constraints on the actual level of public services provided. One view is that government units are able to adjust optimally via a combination of automatic and discretionary responses, to the relative price and real income effects induced by inflation. Here, taxes are a veil and rates simply adjust as needed. Another view is that institutional frictions constrain taxes from fully adjusting in the short run and that the actual real expenditure levels observed at any moment in time may not be equilibrium levels. A "budget effect" can be defined as the difference between the optimal quantity of public services demanded and that which is finally provided. Examples of friction are provided by lagged property value reassessment or the omission of services from the sales tax base. Together with a balanced budget requirement, and the political or legal limits to discretionary rate changes, the jurisdiction will have to settle (temporarily) for a quantity of the public good that is smaller than that desired.

The size of the budget effect will be influenced by two factors. The first is the income elasticity of the tax bases, *ceteris paribus*, a greater effect might be expected from communities relying more heavily on the property tax. The second is the size of the net indebtedness of the community. The budget effect will be smaller if, for example, the community is holding a surplus on which they may draw to reach desired expenditure levels.

III. Empirical Results

Annual data from the National Income and Product Accounts are used here to test for the presence of these three effects over the 1972–88 period. To capture the behavior of real expenditures over the sample period, we consider two definitions of the dependent variable. The first includes all expenditures on goods and services, transfers to individuals and interest payments. The second includes only government purchases. Both measures are deflated by the fixed-weight price index of state and local governments.

Changes in real expenditures over the sample period reflect the impact of inflation, but they also reflect changes in the determinants of real demand which take place independent of inflation. The determinants we take into account here are population and personal income. We control for population growth by measuring the dependent variable, and where appropriate, the explanatory variables, in per capita terms. To control for income changes, we use per capita personal income deflated by the fixed-weight price index for personal consumption as an independent variable.

The explanatory variable for relative prices is measured as the ratio of the fixed-weight price indices for state and local government purchases and for GNP. The real income effect of inflation due to federal income taxes and grants is represented in the regression equations by two separate explanatory variables: per capita federal income taxes deflated by the fixed-weight price index for personal consumption expenditures, and per capita federal grants-in-aid deflated by the fixed-weighted price index for state and local government purchases. The hypothesis is that inflation will erode increases in nominal income because of bracket creep and the federal income tax variable will therefore be negatively related to real expenditures, but that increases in federal grants will stimulate real demand.

To capture the presence of the budget effect, we use the net liability position of the state and local government sector lagged by one year. The variable is measured in per capita terms and deflated by the fixed-weight price index for state and local government purchases of goods and services. Net liability of state and local governments is defined as total debt outstanding minus financial asset holdings in the Flow of Funds Accounts, and in this analysis excludes assets or liabilities of the pension and retirement funds. The hypothesis here is that the budget effect, *ceteris paribus*, will be weaker to the extent the financial asset position is stronger. Because this variable is specified as net liabilities, the coefficient should be negatively signed.

Since the explanatory variables include the relative price, estimation is by two-stage least squares (2SLS) using the state and local government compensation of employees fixed-weight price index and the AAA municipal bond rate as additional instruments. From the different specifications we tried, the log-linear specification performed best, and is reported in Table 1. The Durbin-Watson coefficients do not suggest a serious problem with autocorrelated errors, and a Cochrane-Orcutt adjustment left the estimated equations practically unchanged.

All coefficients take the expected sign and are statistically significant at the 5 percent level. After controlling for population and real personal income changes, the three effects that potentially are caused by inflation, and developed in our model, appear to have played a significant role in the determination of real growth in the state and local government sector in the 1972–88 period.

TABLE 1—2SLS REGRESSION EQUATIONS: REAL PER CAPITA EXPENDITURES FOR THE STATE AND LOCAL SECTOR, 1972–88[a]

Explanatory Variables	Total (1)	Purchases (2)
Relative Price	−1.83	−2.22
	(3.08)[b]	(3.17)[b]
Per capita real personal income	0.95	0.89
	(2.68)[c]	(2.14)[c]
Per capita real federal income tax	−0.44	−0.40
	(4.27)[b]	(3.32)[b]
Per capita real federal aid	0.23	0.30
	(2.84)[c]	(3.19)[b]
Per capita real net liabilities, lagged one year	−0.21	−0.22
	(3.15)[b]	(2.79)[c]
Constant	12.12	11.61
	(2.92)[c]	(2.86)[c]
Adjusted R^2	0.886	0.703
D-W Statistic	2.14	2.12
F-Statistic	25.93	8.57

[a] All variables in logarithms. Absolute value of *t*-statistics in parentheses. Dependent variable: Per capita real expenditures.

[b] Significant at the 0.01 level by a two-tail test.

[c] Significant at the 0.05 level by a two-tail test.

The coefficients for the relative price effect indicate the existence of an elastic response of per capita real expenditures. Previous estimates of price elasticities in the literature have been in the inelastic range, but these estimates were obtained from cross-section, disaggregated data, with nominal (as opposed to real) measures of expenditure, and considered the own-price rather than the relative price effect. The coefficient for personal income shows real demand to be income inelastic although the elasticities are near unity.

The real income effect is as expected in that federal tax liabilities take on a negative sign, and grants are directly related to the level of expenditure. Given the presence of bracket creep and the declining level of real grants over much of the period under study, the real income effect dampened the demand for state and local government expenditures. However, taken at mean values, the estimated marginal propensity to spend on state and local government services is higher for federal grants-in-aid than for personal income. This finding is more consistent with the flypaper effect than the fungibility model assumed here.

The net liability variable has the expected negative sign. When the financial position of the sector is weaker, the observed level of real expenditures is lower. This finding is consistent with our hypothesis about the existence of a "budget effect," due to frictions that slow the adjustment process. It is also consistent with the argument that there is a "wealth effect" that will cause the median voter to demand a higher level of expenditures.

IV. Summary and Conclusions

While the previous literature on role of the inflation in determining the real growth of state and local government budgets has contributed to our understanding of several measurement issues, it has failed to provide an adequate framework to sort out the state and local government response to inflation. Analysts and policymakers alike seem to be confused over whether inflation should be viewed as a fiscal bonus with dominating revenue effects, or as primarily a cost matter and an important contribution to fiscal distress. In this paper we develop a stylized model of state and local behavior and identify three effects of inflation on the real demand for state and local government services: a relative price effect, a real income effect, and a budget effect. The budget effect measures disequilibrium between actual and demanded levels of expenditure, and it should be present only when necessary discretionary adjustments do not take place. In these situations, the budget process is not a veil and has a real impact on demand. Tests for the presence of the three effects, using annual data for state and local governments for 1972–88, give results that are consistent with the model developed here.

REFERENCES

Baumol, William J., "Macroeconomics of Unbalanced Growth: The Anatomy of the Urban Crisis," *American Economic Review*, June 1967, *57*, 415–26.

Beck, Morris, "Public Expenditure, Relative Prices, and Resource Allocation," *Public Finance/Finances Publiques*, No. 1, 1985, 17–33.

Bird, Richard, *The Growth of Public Employment in Canada*, Toronto: Institute for Research on Public Policy/Institute de Recherches Politiques, 1979, ch. 2.

Borcherding, Thomas and Deacon, Robert, "The Demand for Services of Nonfederal Governments," *American Economic Review*, December 1972, *62*, 891–901.

Feenberg, Daniel and Rosen, Harvey, "State Personal Income and Sales Taxes, 1977–1983," in Harvey Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press, 1986.

Greytak, David and Jump, Bernard, "The Impact of Inflation on State and Local Government Finances, 1967–1974," Occasional Paper 25, Metropolitan Studies Program, Maxwell School, Syracuse University, 1975.

Heller, Peter, "Diverging Trends in the Shares of Nominal and Real Government Expenditure in GNP: Implications for Policy," *National Tax Journal*, March 1981, *34*, 61–74.

Ott, Attiat F., "The Effect of Inflation on the Balance Sheet of the State-Local Sector," in George F. Break, ed., *State and Local Finances: The Pressures of the 1980s*, Madison: University of Wisconsin Press, 1983.

Ross, John and Reeder, Richard, *State-Local Finances in Recession and Inflation: An Economic Analysis*, Report No. A-70, Advisory Commission on Intergovernmental Relations, 1979.

State Assistance to Local Governments: Changes During the 1980s

By HELEN F. LADD*

Recent changes in the U.S. federal system have increased the role of subnational governments and focused attention on the interrelationship between state and local governments. This paper analyzes one aspect of this interrelationship, the provision of state aid to local governments. In 1987, state governments devoted on average about one-third of their general spending to intergovernmental aid that in turn accounted for about one-third of all local general revenues and more than 50 percent of the general revenues of independent school districts. These large magnitudes imply that changes in state aid to local governments can have big impacts on the finances of both state and local governments.

Despite the importance of state aid, economists have paid little attention to developing positive empirical models of the state aid decision. (Exceptions are Steven Craig and Robert Inman, 1982; 1986, for education, and John Yinger and I, forthcoming, for central cities.) The extensive economics literature on intergovernmental grants focuses primarily on the behavioral responses of recipient governments and how grants should be designed to achieve various grantor objectives, but it is virtually silent on the grantor's decision process. The standard economic arguments for intergovernmental aid include offsetting interjurisdictional spillovers, offsetting horizontal or vertical disparities, or influencing the distribution of services, such as education, deemed to be meritorious (see George Break, 1980, for a thorough discussion). In principle, one could test whether the economic models appear to have informed state decisions. However, these tests would be difficult to implement because the state aid decision is not independent of how spending responsibilities are allocated between state and local governments, the explanatory variables would be crude (how, for example, would one measure spillovers?), and economic differences across states are likely to be dominated by differences in cultural and historical factors.

By focusing on changes in aid over time, this paper takes the institutional context as given and thereby obviates the need to explain the large variation across states in aid levels. As a consequence, the formal model makes no reference to the magnitude of disparities or spillovers that could be important determinants of variation across states. Instead, the model concentrates on how state governments adjust their intergovernmental aid in response to external shocks such as changes in the state economy, federal aid, and local taxes. The estimated responses should be useful for policy discussions about state aid and the desirability of local revenue diversification.

I. A Model of the State Aid Decision

In developing a positive model of state aid to local governments, the first task is to specify the state policymakers' welfare function. Clearly, state policymakers value the public services (such as highways and higher education) that they provide directly using state government employees. The more difficult issue is how they value services provided by local governments. The starting assumption for this model is that their perception of local services depends on how the local services are financed. Let us assume that the price of local services is $1 and that spending on local services is financed by three revenue sources: locally generated taxes (*LTAX*), aid received from the state (*STAID*), and aid received from the federal government (*FEDAID*). This assumption implies that states perceive the value of local services (*LS*) to be:

---

*Professor of Public Policy Studies, Duke University and Senior Fellow, Lincoln Institute of Land Policy, 26 Trowbridge Street, Cambridge, MA 02138. I thank the Lincoln Institute for support and the National Association of State budget Officers for making their survey data available to me.

$$LS = STAID + a\,LTAX + b\,FEDAID$$

where a and b represent the values that state officials place on dollars of local spending (and hence services) financed by each revenue source relative to the value they place on state-provided services or local services financed with state aid. In general, we would expect the valuation parameter b to be less than one. That is, we would expect state policymakers to value services financed by federal aid less highly than those services financed by state aid.

This relative valuation is most clearly predicted when intergovernmental aid is restricted to narrowly defined categorical programs. Because the programs targeted by federal aid are determined by national, rather than state, policymakers, their value to state policymakers is likely to fall short of that for programs that the state has explicitly chosen. However, the more unrestricted are the aid programs, the less difference we would expect in the two valuation parameters. For example, state policymakers might value equally those local services financed by federal general revenue sharing and those financed by a state revenue-sharing program.

The magnitude of the valuation parameter a is hard to predict. If state policymakers care little about the service level provided by local governments out of their own revenues, a would be close to zero. Alternatively, if state policymakers view locally financed services as perfect substitutes for services they provide directly, a would have the value 1. But there is more to the story. A value of 1 on locally financed services would imply that states would have incentives to shift more and more responsibilities onto local governments since locally provided services would not consume scarce state budgetary resources. Hence, to make the model more realistic, state policymakers are assumed to care in addition about the local tax burden; a higher local tax burden decreases the welfare of local taxpayers who in turn may complain to state policymakers.

Thus, a higher local tax burden has potential political costs and decreases the welfare of state policymakers. This addition leads to the concept of state policymakers' perceived value of net local services (*NLS*), that is, local services net of the local tax burden:

$$NLS = STAID + (a - c)\,LTAX + b\,FEDAID$$

where c (defined as a positive number) represents the perceived political cost to state policymakers of local taxes. The expression $(a - c)$ represents the net value to state officials of local services financed by local taxes. If a = c, state officials are indifferent to the level of services that local governments provide out of their own resources.

State policymakers are assumed to allocate a fixed level of budgetary resources (B) between state aid (*STAID*) and direct state services (G) in order to maximize a welfare function whose arguments are direct state services and the perceived level of local services net of taxes. Assuming for simplicity that state policymakers' preferences are defined by a Cobb-Douglas welfare function, the state policymaker's problem is to

$$\text{Max}\; W = G^{\alpha} NLS^{\beta}$$

subject to $B = G + STAID$

$$NLS = STAID + (a - c)\,LTAX + b\,FEDAID$$

$$LTAX = f(STAID, X)$$

where X is a vector of exogenous determinants of local taxes and $\delta LTAX/\delta STAID \leq 0$. According to this formulation, the "price" to the state of purchasing more net local spending through state aid depends on $(a - c)$. If the value that state policymakers place on locally financed public services is fully offset by the political cost to the state of the local taxes used to finance those services, so that $(a - c)$ equals 0, the price to the state of obtaining additional net local services is \$1.00. If $(a - c)$ exceeds zero, then the price generally exceeds \$1.00 and if

$(a - c)$ is less than zero the price is less than \$1.00.

As elaborated more fully in my 1990 paper, solving this model and making the variables operational yields an equation in which per capita changes in state aid are explained as a linear function of changes in various components of the state budget including state own-source revenues (*SOSREV*) and federal aid to states for social services (*FEDSTSS*) and for other purposes (*FEDSTOTH*); local own-source revenues (*LOSREV*); and direct federal aid to localities for social services (*FAIDSS*) and for other purposes (*FAIDOTH*). In addition, a variable representing general fund balances at the beginning of the period (*BAL*82) is included to control for differences across states in initial fiscal condition.

II. Results

The model is estimated using two-stage least squares for real per capita changes between 1982 and 1987 for 44 states; 6 states were excluded for reasons of atypical fiscal structure, small size, or missing data. State and local own-source revenues (*SOSREV* and *LOSREV*) are both treated as endogenous variables.[1] Table 1 reports parameter estimates for four forms of the dependent variable.

The dependent variable in the preferred equation in column 1 is changes in state aid as measured by survey information from the National Association of State Budget Officers (NASBO). This measure labeled *NSTAID* is preferable to the comparable measure, *CSTAID*, in column 2 (based on data from the Census of Governments) in that it excludes from state aid, funds from the federal government that are simply passed through the state to local governments. Both measures of changes in state aid have been adjusted upward to include the monetary value of state takeover from local governments of services such as courts and jails during the period using data reported by NASBO. The dependent variables in columns 3 and 4 represent changes in unrestricted or general aid. These are designated by *NSTAIDU* for unrestricted aid as reported by NASBO, and *CSTAIDGEN* for general purpose aid as reported by the census.

TABLE 1—STATE AID TO LOCAL GOVERNMENTS
(PER CAPITA CHANGES, 1982–87)

	(1)	(2)	(3)	(4)
Endogenous Variables				
SOSREV	0.296^{a}	0.217^{a}	0.044^{b}	0.038
	(4.23)	(2.87)	(1.60)	(1.27)
LOSREV	0.114	0.066	−0.032	−0.030
	(0.83)	(0.44)	(0.58)	(0.51)
Exogenous Variables				
FAIDOTH	−0.979^{a}	−0.951^{a}	−0.240	−0.380^{a}
	(2.06)	(1.86)	(1.29)	(1.89)
FAIDSS	−0.166	1.812	0.735	0.155
	(0.10)	(1.01)	(1.13)	(0.22)
FEDSTOTH	0.152	0.214	−0.044	−0.119
	(0.39)	(0.51)	(0.29)	(0.72)
FEDSTSS	0.424^{b}	1.812	−0.082	−0.098
	(1.38)	(1.02)	(0.68)	(0.744)
*BAL*82	167.22^{b}	161.27	45.83	72.99^{b}
	(1.34)	(1.20)	(0.93)	(1.38)
Constant	−54.48^{b}	−46.55	−1.67	−8.89
	(1.56)	(1.24)	(0.12)	(0.60)
R^2	.54	.29	.17	.23
Mean Square Error	49.64	53.48	19.49	21.15
F-Statistic	5.93	2.10	1.04	1.51
	($p = 0.00$)	($p = 0.04$)	($p = 0.40$)	($p = 0.16$)

Notes: Estimated by two-stage least squares; absolute value of *t*-statistics are shown in parentheses. The dependent variables by equation are (1) *NSTAID*, (2) *CSTAID*, (3) *NSTAIDU*, and (4) *CSTAIDGEN*. Their means and standard deviations (in parentheses), in order, are \$60.49 (\$66.65), \$40.37 (\$58.08), \$8.37 (\$19.56), and \$2.58 (\$22.02). Additional descriptive data are available in my earlier paper.

[a,b]Statistically significant at the 5 and 10 percent one-tailed levels, respectively.

The results in column 1 are reasonable and consistent with the theoretical model. In particular, the main measure of state budgetary resources (*SOSREV*) has a positive impact on state aid; the major direct federal aid variable (*FAIDOTH*) enters with a negative sign; and changes in the locally generated revenue variable (*LOSREV*), whose predicted coefficient depends on the relative values of locally financed services and local taxes to state policymakers (the value of $(a-c)$), has a coefficient that is statistically indistinguishable from zero. In brief, these results imply that state budgetary resources affect the local aid decision, that states almost fully offset changes in nonsocial services aid to local governments, and that the value to state policymakers of additional locally financed public services is balanced by the political costs of higher local taxes.

---

[1] The following variables were used as excluded instruments to identify the *SOSREV* and *LOSREV* variables: changes in gross state product and the state unemployment rate as measures of the condition of the state economy; changes in population, Medicaid recipients, and prisoners as measures of state expenditure pressures; changes in per capita income and public school enrollments as measures of the demand for local services; a dummy variable for the existence of stringent limits on local taxes or spending; and a dummy variable for expanded use of local nonproperty taxes.

The results in column 2 based on the census definition of state aid are quite similar to those in column 1. Some minor differences are noted below. In general, the equations for unrestricted or general aid in columns 3 and 4 have less explanatory power than those for total aid, but yield signs and magnitudes that, given the smaller quantities of aid, are roughly comparable to those for total aid.

III. Policy Implications and Discussion

From a policy perspective, the most interesting results are the responses to changes in direct federal aid to localities and to state budgetary resources.

State Response to Changes in Federal Aid to Localities. The estimated response of total state aid to changes in direct federal aid to local governments is striking. The $-.98$ coefficient for *FAIDOTH* implies that between 1982 and 1987, state governments on average offset each $1 change in federal aid (other than that for education and public welfare) to local governments by changing state aid by 98 cents in the other direction. This represents a larger response than that found in previous studies. The results in column 2 confirm that this finding is not a quirk of the NASBO data; even the census variable yields a 95 cent offset. Columns 3 and 4 show that between 24 and 38 cents of the offset is the form of unrestricted or general aid.

Several points about this finding should be noted. First, it should not be interpreted as the state response to the 1986 elimination of general revenue sharing. Instead, additional tests (reported in my 1990 paper) show that it represents the response to changes in a variety of other federal programs throughout the 1982–87 period. Additional tests show that the finding cannot be explained away either by the fact that the variables are absolute rather than percentage changes, or that states were in disequilibrium with respect to their adjustment to cuts in federal aid prior to 1982. After all appropriate refinements and tests, the state aid offset remains larger than 80 cents per $1 change in federal aid.

The findings differ for the other direct federal aid variable. The small negative and insignificant coefficient of *FAIDSS* in the *NSTAID* regression implies that between 1982 and 1987, state governments essentially offset none of the change in direct federal aid to local governments for education and public welfare. In terms of the theoretical model, this finding is consistent with state officials placing a low value on certain locally provided services that are financed by the federal government. The larger coefficient for this variable in the *CSTAID* equation probably reflects the inclusion of federal pass-through aid in the *CSTAID* variable; in states where local governments take on substantial responsibility for social services, local governments are likely to receive above-average amounts of direct federal aid for social services as well as above-average amounts of federal pass-through funds.

State Response to Changes in Budgetary Resources. Both total and unrestricted aid vary positively with exogenous growth in state-own source revenues. For each $1 increase in state revenues, total state aid increases by about 30 cents using the NASBO definition and about 22 cents using the census definition. The smaller coefficient for the census variable is explained by the fact that the federal pass-through funds that are included in the census variable are less responsive to the budgetary condition of the state than is state aid financed from own sources.

Translating the estimated coefficients into response elasticities shows that the elasticity for total aid with respect to changes in the state budget is less than 1. In particular, a 10 percent increase in state own-source revenues leads to only a 7.7 percent increase in state aid (as defined by NASBO). In contrast, the elasticities for unrestricted or general aid are close to one. These higher elasticities emerge because much of this unrestricted aid takes the form of shared state tax revenues that are expected to grow at the same rate as other state tax revenues. The lesson for local officials is that they are more likely to share in the growth of the state economy when aid is in the form of shared taxes than when it is dependent on state appropriations.

Another component of state budgetary resources is federal aid to the state governments. The coefficients on one of the federal-to-state aid variables (*FEDSTOTH*) are small and insignificant. These results suggest that the federal government should not expect state governments to act as its agents in assisting local governments to provide general public services. According to the first regression, a more optimistic conclusion emerges with respect to federal aid for social services such as welfare and education, but this result deserves further investigation.

Choice of Period. Dictated by the availability of the NASBO data, the 1982–87 period for this analysis is advantageous because it encompasses a period of large changes in federal aid to state and local governments. Unfortunately, 1982 was not an equilibrium year for states. Not only had they experienced unprecedented cuts in federal aid in 1981 and 1982, but also the economy was in deep recession. But how that disequilibrium affects the behavioral estimates reported here is unclear. Moreover, the 1982 general fund balance variable (*BAL*82) helps control for differences across states in their 1982 fiscal condition, thereby minimizing potential biases associated with variation in the degree of fiscal disequilibrium across states. The positive sign on this variable suggests that the stronger is a state's fiscal condition at the beginning of the period, the more additional assistance it is able to give to its local governments in following years. This finding provides additional support for the view that when local governments rely heavily on state aid, they become vulnerable to the vagaries of state government fiscal problems.

IV. Conclusion

As an important component of state-local relations, state aid to local governments deserves more attention from political economists. Building on the work of Yinger and myself, this paper specifies a state government objective function and budget constraint that yield estimable models of the state aid decision. Central to the objective function is how state policymakers value services provided by local governments. Further research on this valuation issue would be fruitful. With respect to policy, states throughout the country are struggling to determine the appropriate mix of state aid and locally generated revenues. The results in this paper suggest that too much reliance on state aid could be harmful to the fiscal health of local governments.

REFERENCES

Break, George, *Financing Government in a Federal System*, Washington: Brookings, 1980.

Craig, Steven G. and Inman, Robert P., "Federal Aid and Public Education: An Empirical Look at the New Fiscal Federalism," *Review of Economics and Statistics*, November 1982, *64*, 541–552.

______ **and** ______, "Federalism, Welfare, and the "New Federalism": State Budgeting in a Federalist Public Economy," in Harvey S. Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press, 1986.

Ladd, Helen F., "State Aid to Local Governments in the 1980s," Lincoln Institute of Land Policy, 1990.

Yinger, John and Ladd, Helen F., "The Determinants of State Assistance to Central Cities," *National Tax Journal*, forthcoming.

Commercial Land Use Regulation and Local Government Finance

By ERIC A. HANUSHEK AND JOHN M. QUIGLEY*

Land use regulation in American cities is pervasive. Taken together, these rules probably represent the most significant market intervention undertaken by state and local governments. Land use regulation includes a wide variety of rules governing the physical location of economic activities within jurisdictions; regulations governing the design, height, or capital intensity of commercial and industrial property; rules regarding the minimum lot sizes for residential parcels, the number of bedrooms in new dwellings or other restrictions on residential density; and rules delineating developer responsibilities for infrastructure provided in newly subdivided areas.

By international standards, regulations in the United States are adopted by rather small units of government (i.e., by towns rather than regional authorities) whose objectives are typically confined to narrowly defined geographical portions of metropolitan areas. The small scale at which regulations are promulgated, and the existence of substitutable sites within metropolitan areas for residential, commercial, and industrial activity, suggest that recognition of competition across jurisdictions is crucial to understanding the impact of land use rules as they affect the quality of life or the fiscal health of the residents of any jurisdiction. Many economic models ignore these competitive forces, concentrating instead upon the case of the regulatory monopolist. In consequence, there is considerable theoretical uncertainty about the impact of locally adopted regulations upon the utilities of local citizens or their economic wealth. (A selective review is provided by J. M. Pogodzinski and Tim Sass, 1989.)

Land use regulations in urban areas are motivated by two conceptually distinct concerns: the control of externalities and the pursuit of fiscal objectives. These mixed objectives make it more difficult to understand the effects of regulation, especially since externality arguments are notoriously difficult to confirm in empirical analyses. Fiscal arguments, on the other hand, are more easily tested.

In this paper we investigate a new but increasingly popular form of land use control: limitations on the growth of commercial property. In one instance, we provide bounds on the fiscal implications of controls on commercial property; this, in turn, bounds the magnitude of the implied efficiency gains when the land use policy is viewed as a rational choice by citizens.

I. Externality and Fiscal Zoning

The classic efficiency rationale for adopting zoning regulations is to mitigate the adverse physical consequences that some economic activity may hold for those activities conducted on neighboring sites. This argument, and the vivid references to smokestacks and laundries, goes back at least to Pigou and later to Coase; it forms the basis for rules segregating industry from residential areas. Closely related to the physical externalities regulated by zoning is the possibility of using these regulations to control social externalities, as when members of one socioeconomic group are considered more "desirable" as neighbors than are members of other socioeconomic groups. Both physical and social externalities can be modeled in a similar fashion, and the results of such models indicate the narrow static efficiency of segregation of households or land uses.

One implication of the efficiency of externality zoning is its effect upon aggregate land prices. By internalizing the external ef-

---

*University of Rochester, Rochester, NY 14627, and University of California, Berkeley, CA 94720, respectively.

fects associated with land use, or merely by minimizing the border between incompatible land uses, these zoning regulations increase the aggregate value of land within the community, and hence they increase the local tax base.

Fiscal zoning regulations, in contrast, are intended to increase the profitability of the local public sector directly by internalizing the pecuniary externalities that arise through the tax and expenditure system from different uses of urban land. These motives are manifest in regulations encouraging the location of those firms that produce more in tax revenues than they use in public expenditures on services. Given the reliance of local government upon property tax revenues, fiscal zoning in the residential sector typically involves the adoption of rules requiring households using public services to consume (and to pay taxes on) more real property than they would otherwise choose.

In contrast to externality zoning, it need not follow that fiscal zoning increases aggregate land prices, and such zoning certainly need not imply aggregate welfare gains. Indeed, the typical residential zoning regulation specifying a minimum lot size lowers the value of land held for residential development by eliminating the demand for smaller parcels. The profitability of fiscal zoning for a local community depends upon a comparison of tax base changes and variations in the expenditures for public services which can be attributed to the regulations.

II. Observable Effects of Zoning

As an empirical matter, it is extremely hard to sort out the pecuniary from the externality motives for zoning, and equally hard to distinguish their separate effects. Empirical evidence generally supports the conclusion that restrictive land use controls raise housing prices in communities that impose them (William Fischel, 1989), but there is little evidence that could be used to distinguish the sources of these price increases. Reductions in supply arising from zoning regulation could lead to price increases for housing as long as the demand for sites were not perfectly elastic. However, increases in amenities arising from the control of externalities could also lead to housing price increases in response to land use restrictions.[1] Existing empirical evidence also suggests that values for vacant or agricultural land are lower at locations in proximity to jurisdictions imposing more stringent controls (Jan Bruckner, 1989).

In virtually all of these empirical analyses, zoning refers to land use restrictions in the residential sector. Little empirical analysis exists on the effects of commercial or industrial zoning on municipal revenues, and studies of the effects of externalities on housing prices have seldom included measures of congestion or agglomeration, which are presumably affected by land use regulation in the nonresidential sector.[2] (This literature is reviewed by Timothy Bartik and V. Kerry Smith, 1987.)

III. Nonresidential Zoning

During the past decade, there have been ever-more vigorous attempts to limit nonresidential development to increase fiscally profitable activities and to improve urban amenities. (See Robert Burchell and David Listoken, 1978, for a variety of methods of estimating "profitability.") This movement has been especially strong in California where additional growth is perceived to entail high environmental costs, where average incomes and the demand for urban amenity is high, where ad valorem tax rates are low (and essentially uniform), and where it is perceived that the demand for location by firms is inelastic.[3]

---

[1] In this latter case, the increase in housing prices (values) would arise from an increase in the quantity of service provided.

[2] There is, however, a vast literature on the relationship between air pollution and housing prices. Presumably, levels of ambient air quality are sensitive to the locations of firms and households.

[3] Indeed, during the 1971–86 period there were 133 local initiative referenda in the State of California on proposals for nonresidential growth control, building restrictions and moritoria, and so forth. In San Francisco alone, 7 such initiatives appeared on the ballot (see California Association of Realtors, 1987).

The effect of nonresidential zoning on local public finance depends, on the one hand, on the effectiveness of regulation in enticing or repelling tax-paying businesses and, on the other hand, in the net change in public resources devoted to servicing those businesses, their tenants, customers, and employees. More generally, the effect of these regulations on citizen welfare depends also on any amenity changes associated with the regulations—variations in congestion, pollution and urban asthetics.

The welfare effects of these proposals may be quite difficult to measure, but the fiscal effects of limitations on commercial development are somewhat less elusive. We consider, for example, the fiscal effects of one dramatic case from urban California, limitations on commercial development adopted by initiative in San Francisco in 1986 (Proposition M, approved by San Francisco voters by a 50.6 percent vote). This initiative established an annual limit (950,000 sq. ft.) on new office space and regulated its distribution among buildings of varying sizes.

The direct fiscal effects of this land use restriction arise from changes in the value of commercial property, changes in the value of residential property, and from changes in service requirements and government expenditures. The key elements entering into property valuations are the effects of the restriction on commercial and industrial rents, and upon the level of employment. The capitalized stream of rents to properties provides a direct indication of property values subject to local tax. The employment patterns provide evidence on the demand for residential housing, since reduced employment will translate directly into reduced demand for city housing.

A complete analysis of rents, employment, and location decisions would require a structural model of the urban area, but there are more limited models that can be used to project the rent and employment effects. We employ two attempts to model the market for commercial property (Kenneth Rosen and Ruth Shragowitz, 1985, and Bradford Case and Deborah Gordon, 1989) to bound the incremental effects of commercial growth restrictions. These studies provide time-series estimates of commercial employment and office space, new construction, and rents in the San Francisco area. We use these models to compare rents and employment if the economy evolved under this restriction with projections made in the absence of this control over land use.

TABLE 1—EFFECTS OF PROPOSITION M ON SAN FRANCISCO EMPLOYMENT GROWTH AND COMMERCIAL RENTS[a]

Model	Annual Employment Growth	Annual Rent Increases
Rosen-Shragowitz	−0.9	+3.9
Case-Gordon	−0.8	−4.9

[a] Shown in percent.

Table 1 displays our projections of differences in employment growth and commercial rent increases over the first ten years of growth controls. Both projections indicate that the growth limitation measure will have a negative effect on employment, reducing growth in office employment by approximately 1 percent per year. They diverge, however, in their estimates of the effect on rents. The estimates based on the Case-Gordon (C-G) model suggest that rents would actually be less in the constrained case; the difference in annual increase in rents, other things being equal, would amount to almost 5 percent per year. The Rosen-Shragowitz (R-S) model, on the other hand, supports the conclusion that the growth restrictions would increase office rents. The C-G analysis provides an unambiguous forecast of fiscal losses, arising from reductions in the values of both commercial and residential property. The R-S analysis, however, leaves some ambiguity, since the rise in rents is greater in the restricted case (by 3.9 percent per year). Moreover, the aggregate reductions in total office space attributable to Proposition M are less than the aggregate rent difference over the ten-year period, implying that total commercial property values increase.[4] By these estimates, the total fiscal effects balance residential losses against commercial gains.

---

[4] The R-S model projects that growth in central city office space with growth controls will be 18 percent less

The simulation results underscore the fact that actions taken by one jurisdiction within the metropolitan area lead to mobility of both households and firms, and this alters property values and the tax base.

How can these growth restrictions be rational choices by the citizens of San Francisco? The employment changes suggest that, *ceteris paribus*, demand for residential property will fall, leading to capital losses and lower fiscal capacity from that source. If commercial rents also fall (as projected by the C-G model), there is an unambiguous decline in commercial property values. Fiscal gains could then result only if the new commercial development excluded under the limitation demanded far more in services than its tax payments—enough to make up for the loss in both commercial and residential tax bases. Even with fiscal losses, the policy could still be rational, but only if the efficiency gains from controlling externalities were quite large, given the projections reported in Table 1.

If commercial rents increase (as projected by the R-S model), the picture changes somewhat. Here, the current owners of commercial property get windfall gains, and the owners of residential properties suffer windfall losses from reduced locational demands. Homeowner losses from changed employment patterns may be offset by increased net tax revenues from commercial properties (fiscal effects), and by externality reductions that might separately be capitalized into values. It is nevertheless worth noting that the clearest gains accrue to the current owners of commercial property and not to the residents who voted the policy into effect.

The overall results, while not decisive, do raise questions about the underlying rationale for new restrictions on commercial properties. They are likely to be costly on narrow fiscal grounds. Moreover, since these restrictions apply to office space and would generally restrict white-collar employment, any externality effects presumably arise primarily from reduced congestion. These is little reason to believe, however, that growth restrictions are the least cost way of dealing with congestion, even when such externalities are sufficient to warrant direct intervention.

The fiscal effects identified here may be dampened by the sluggish response of the property tax base under California's Proposition 13, and this form of zoning may confer rents on subsets of property owners or employees. Nevertheless, rent seeking does not provide a really satisfactory explanation of the motivations underlying approval of the referendum.

IV. Conclusions

Zoning and land use policies aimed at residential properties generally can be explained as programs to enhance the property values of existing residents. This result, generally confirmed by empirical analyses, arises from combination of efficiency and fiscal effects. Policies toward commercial properties, particularly those limiting growth, are, however, much more difficult to understand from the standpoint of rational voter behavior. Since these policies reduce employment and therefore reduce housing demands, the gains must come from offsetting fiscal or efficiency factors. On the fiscal side, our limited empirical investigation questions the rational basis for these policies.

REFERENCES

Bartik, Timothy J. and Smith, V. Kerry, "Urban Amenities and Public Policy," in Edwin S. Mills, ed., *Handbook of Regional and Urban Economics*, Vol II, Amsterdam: North-Holland, 1987.

Bruckner, Jan, "Growth Controls and Land Values in an Open City," BEPR Faculty Working Paper No. 89-1594, University of Illinois-Urbana-Champaign, August 1989.

Burchell, Robert W. and Listokin, David, *The Fiscal Impact Handbook*, New Brunswick:

---

than that projected without controls by 1999. The G-S model implies even larger space responses, amounting to a 22 percent net decline arising from growth controls.

Center for Urban Policy Research, 1978.

Case, Bradford and Gordon, Deborah, "San Francisco's Office Growth Limit, A Report to the Mayor of San Francisco," mimeo., University of California-Berkeley, January 1989.

Fischel, William, "Do Growth Controls Matter: A Review of Empirical Evidence on the Effectiveness and Efficiency of Local Government Land Use Regulation," mimeo., Dartmouth College, 1989.

Pogodzinski, J. M. and Sass, Tim R., "The Theory of Zoning: A Critical Review," mimeo., San Jose State University, August 1989.

Rosen, Kenneth T. and Shragowitz, Ruth, "The Proposed Growth Limit on Commercial Construction for San Francisco," Working Paper 85-94, Center for Real Estate and Urban Economics, University of California-Berkeley, May 1985.

California Association of Realtors, "Matrix of Land Use Planning Measures, 1971–1986," mimeo., 1987.

THE RATIONALITY OF THE FOREIGN EXCHANGE RATE[†]

Chartists, Fundamentalists, and Trading in the Foreign Exchange Market

By JEFFREY A. FRANKEL AND KENNETH A. FROOT*

The overshooting theory of exchange rates seems ideally designed to explain some important aspects of the movement of the dollar in recent years. Over the period 1981–84, for example, when real interest rates in the United States rose above those of its trading partners (presumably due to shifts in the monetary/fiscal policy mix), the dollar appreciated strongly. It was the higher rates of return that made U.S. assets more attractive to international investors and caused the dollar to appreciate. The overshooting theory would say that, as of 1984 for example, the value of the dollar was so far above its long-run equilibrium that expectations of future depreciation were sufficient to offset the higher nominal interest rate in the minds of international investors. Figure 1 shows the correlation of the real interest differential with the real value of the dollar, since exchange rates began to float in 1973.

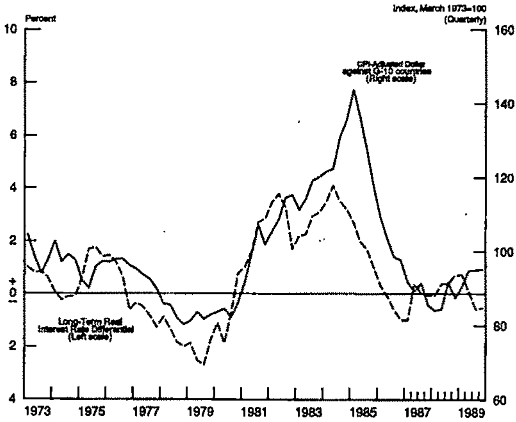


FIGURE 1.
THE DOLLAR AND REAL INTEREST RATES

Source: Peter Hooper and Catherine Mann, Federal Reserve Board.
Notes: Quarterly data. The CPI-adjusted dollar is a weighted-average index of the exchange value of the dollar against the currencies of the foreign G-10 countries plus Switzerland, where nominal exchange rates are multiplied by relative levels of CPIs. Weights are proportional to each foreign country's share in world exports plus imports from 1978 through 1983. The long-term real interest differential is the U.S. rate minus the weighted average of foreign-country rates.

I. Bubble Episodes

At times, the path of the dollar has departed from what would be expected on the basis of macroeconomic fundamentals. The most dramatic episode is the period from June 1984 to February 1985. The dollar appreciated another 20 percent over this interval, even though the real interest differential had already begun to fall. The other observable factors that are suggested in standard macroeconomic models (money growth rates, real growth rates, the trade deficit) at this time were also moving in the wrong direction to explain the dollar rise.

It is now widely accepted that standard observable macroeconomic variables are not capable of explaining, much less predicting *ex ante*, the majority of short-term changes in the exchange rate. But economists divide into two camps on what this means. One view is that the unexplained short-term changes must be rational revisions in the market's perception of the equilibrium exchange rate due to shifts in "tastes and technologies," even if the shifts are not observable to macroeconomists in the form of standard measurable fundamentals. A major difficulty with this interpretation is that it is difficult to believe that there could have been an increase in the world demand for U.S. goods (or in U.S. productivity) sufficient to increase the equilibrium real exchange rate by more than 20 percent over a 9-month period, and that such a shift would then be reversed over the subsequent 9 months.

[†] *Discussants*: Michael Gavin, Columbia University; Karen K. Lewis, New York University; Kenneth D. West, University of Wisconsin-Madison.

*Department of Economics, University of California, Berkeley, CA 94720, and Graduate School of Business, Harvard University, Boston MA 02163, respectively. We thank MMS International for data, and Lu Cordova and Joseph Mullally for valuable research assistance.

This brings us to the second view: that the appreciation may have been an example of a speculative bubble—that it was not determined by fundamentals, but rather was the outcome of self-confirming market expectations. The dollar in 1985 "overshot the overshooting equilibrium." Some have suggested that the appreciation of 1988–89, on a smaller scale, may also have been of this nature.

There exist elegant theories of *rational* speculative bubbles, in which all participants know the correct model. Some observers have suggested that 1984-85 may be best described as a bubble that was *not* characterized by rational expectations.[1] We have suggested earlier that such episodes may best be described by models of bubbles in which market participants do not agree on the model for forecasting the exchange rate (see our forthcoming paper).

While the conventional approach in the literature, theoretical as well as empirical, is to assume that there is such a thing as "the" market expectation of the future exchange rate, there is evidence that investors have heterogeneous expectations. For one thing, surveys of the forecasts of participants in the foreign exchange market show wide dispersion at any point in time. One, conducted by the *Financial Report* (affiliated with the *Economist*), reports a high-low range of 6-month forecasts that averages 15.2 percent. Data in a survey conducted by MMS International show a dispersion of opinion (as measured by the standard deviation across respondents) at the 1-month horizon that averaged 2.2 percent for the yen/dollar rate. The dispersion was slightly higher for the mark, pound, and Swiss franc rates.

II. Trading in the Foreign Exchange Market

The tremendous volume of foreign exchange trading is another piece of evidence that reinforces the idea of heterogeneous expectations, since it takes differences among market participants to explain why they trade. The Federal Reserve Bank of New York has released its 3-yearly count of transactions in the U.S. foreign exchange market. It showed that in April 1989, foreign exchange trading (adjusted for double-counting) totaled \$128.9 billion a day, an increase of 120 percent from March 1986. Simultaneous counts in London and Tokyo reported \$187 billion and \$115 billion a day, respectively. Thus the worldwide total is over \$430 billion of foreign exchange trading a day.

Interestingly, the banks in the New York Federal Reserve Bank census reported that only 4.9 percent of their trading was with a nonfinancial firm, and the nonbanks only 4.4 percent; in other words, 95 percent of the trading takes place among the banks and other financial firms, rather than with customers such as importers and exporters. Clearly, trading among themselves is a major economic activity for banks.

What is the importance of trading volume (beyond motivating the importance of heterogeneous expectations)? There are three possible hypotheses, with regard to implications for movements in the market price. 1) The higher the liquidity or "depth" of the markets, the more efficiently is news regarding economic fundamentals processed and the smaller is "unnecessary volatility" in the exchange rate. 2) The foreign exchange market is already perfectly efficient, so that trading volume is irrelevant to price movements and therefore uninteresting. 3) Much trading is based on "noise" rather than "news," and leads to excessive volatility.

---

[1] Paul Krugman (1985) was one of the first to suggest that the market did not appear to realize the extent to which the appreciation of the dollar was not sustainable. Charles Engel and James Hamilton (forthcoming) find that long-term swings are a general characteristic of exchange rates, and that they are not adequately reflected in the forward market. Such findings of predictable excess returns are standardly interpreted as risk premiums. But evidence from survey data on expectations of market participants suggests that the prediction errors of the forward market are *not* due to risk premiums (see our 1989 paper).

Choosing convincingly among these three hypotheses may be too large a task to accomplish here. But there is evidence that trading volume, exchange rate volatility, and the dispersion of expectations among forecasters are all positively related. We have recently developed a weekly data set for four currencies (British pound, German mark, Japanese yen, and Swiss franc), covering the period October 1984 to February 1988. Trading volume is measured by the weekly number of futures contracts (nearest-term) traded on the International Monetary Market of the Chicago Mercantile Exchange; volatility is measured by the squared percentage 15-minute changes in the futures price, averaged over the week; and dispersion is measured by the percentage standard deviation of forecasts across respondents in the survey conducted weekly by MMS International.

Granger-causality tests on prewhitened data show that the degree of dispersion has strong effects on the market. Dispersion Granger-causes volume at the 95 percent level in three currencies out of four, and dispersion also Granger-causes volatility. We also find that the contemporaneous correlation between volume and volatility is high.

One interpretation of these results is that the existence of conflicting forecasts leads to noise-trading (the causation runs from dispersion to the volume of trading, and then from trading to volatility), though there probably exist other interpretations as well. (The Granger test does not show statistically significant causation running directly from volume to volatility. But one would expect any such causality to be purely contemporaneous, and it is important to keep in mind that the Granger test cannot detect this type of causality.)[2]

III. The Rising Importance of Chartists

We now turn to the question of how the existence of different forecasting techniques might lead to "excess volatility." It has long been remarked that if there exist traders who tend to forecast by extrapolating recent trends (i.e., who have "bandwagon expectations"), then their actions can exacerbate swings in the exchange rate. Many so-called "chartist" forecasters, or technical analysts, are thought to use rules that are extrapolative, such as, "Buy when the 1-week moving average crosses above the 12-week moving average."

How do speculators form expectations in practice? Our forthcoming paper offers evidence from the survey data that, at short horizons, respondents tend to forecast by extrapolating recent trends, while at long horizons they tend to forecast a return to a long-run equilibrium such as purchasing power parity. Table 1 reports an update of these estimates. The coefficients reported are to be interpreted as answers to the question, "for every 1 percent that the dollar appreciates in a given week, what percentage change does the median respondent forecast for the dollar thereafter?" The answer at the 1-week horizon is another .13 percent in the same direction. At the 4-week horizon, the extrapolation is smaller. Respondents expect that, by the time 3 months have passed, the dollar will be *lower* than at the day when they are formulating their forecasts, and lower still at 6 months. One year out, they expect the dollar to be .33 percent lower, for every 1 percent that the dollar has appreciated this week.

This leads to the question: which kind of forecasters dominate the market, those who

---

[2] It should be noted that the tests also show that volatility Granger-causes dispersion. We think that this apparent effect may be partly spurious: the MMS survey catches different respondents at different times of the day, so their forecasts of the expected future level of the exchange rate will differ more if the level of the spot rate on that day moves around more. It is also possible that higher lagged volatility causes higher dispersion of expectations because forecasters use different models to interpret the data. All results are reported in our 1990 working paper.

TABLE 1—DO FORECASTERS EXTRAPOLATE?

Survey Data Source and Sample Period	Term of Forecast	Estimate of Extrapolative Parameter	t-Ratio (with GMM standard error)
MMS International Oct. 1984–Jan. 1988	1 week	.13	4.32[a]
	4 week	.08	1.60
Economist June 1981–Aug. 1988	3 month	−.08	−2.98[a]
	6 month	−.17	−4.98[a]
	12 month	−.33	−5.59[a]

Note: OLS Regressions of expected future rate of depreciation against most recent actual depreciation.

[a] Significant at 99 percent confidence level.

TABLE 2—TECHNIQUES USED BY FORECASTING SERVICES

Year	Total	Chartist	Fund.	Both
1978	23	3	19	0
1981	13	1	11	0
1983	11	8	1	1
1984	13	9	0	2
1985	24	15	5	3
1988	31	18	7	6

Source: Euromoney, August issues. *Notes:* Total = number of services surveyed; Chartist = number who reported using technical analysis; Fund. = number who reported using fundamentals models; and Both = number reporting a combination of the two. When a forecasting firm offers more than one service, each is counted separately.

think short term, and appear to have bandwagon expectations, or those who think long term and have regressive expectations? Since Milton Friedman (1953), the standard argument against the importance of destabilizing speculators is that they will on average lose money, and be driven out of the market in the long run. A number of special counterexamples to the Friedman argument have been constructed over the years, most involving heterogeneous actors (for example, "suckers" who lose money and "sharpies" who win). The simplest counterexample would be based on the theory of rational speculative bubbles, where each market participant loses money if he *does not* go along with the herd. The problem with this theory, which identifies speculative bubbles with the unstable paths in a rational-expectations saddle-path problem, is that it has nothing to say about what causes a bubble to start. What, for example, generated a speculative bubble in the period leading up to February 1985, if that is what the dollar surge evident in Figure 1 was?

The model of speculative bubbles developed in our forthcoming paper says that, over the period 1981–85, the market shifted weight away from the fundamentalists, and toward the technical analysts or "chartists." This shift was a natural Bayesian response to the inferior forecasting record of the former group, as their forecasts of dollar depreciation continued to be proven wrong month after month. The change in the weighted-average forecast of future dollar depreciation in turn increased the demand for dollars, and therefore its price in the foreign exchange market.

Is there any sort of evidence for such a theory? *Euromoney* magazine runs a yearly August review of between 10 and 27 foreign exchange forecasting firms. Summary statistics are reported in Table 2. The trend is very clear. In 1978, 18 forecasting firms described themselves as relying exclusively on economic fundamentals, and only 2 on technical analysis. By 1985, the positions had been reversed: only 1 firm reported relying exclusively on fundamentals, and 12 on technical analysis.[3]

In short, it may indeed be the case that shifts over time in the weight that is given to different forecasting techniques are a source of changes in the demand for dollars, and that large exchange rate movements may take place with little basis in macroeconomic fundamentals.

---

[3] A number of firms combine the two approaches, or else offer a separate service of each kind; in this case, usually technical analysis is used for short-term forecasting and fundamentals for long-term forecasting. This pattern matches up well with the regression results from surveys of market participants regarding exchange rate expectations, reported above. The pattern is also confirmed in Helen Allen and Mark Taylor (1989, p. 4), who report that, at short horizons, approximately 90 percent of respondents use some chartist input in forming their expectations, and 60 percent judge charts to be as important as fundamentals, while at the horizon of one year and longer, nearly 30 percent rely purely on fundamentals, and 85 percent judge fundamentals to be more important than charts.

REFERENCES

Allen, Helen, and Taylor, Mark, "Chartists, Noise and Fundamentals: A Study of the London Foreign Exchange Market," CEPR Discussion Paper No. 341, September, 1989.

Engel, Charles and Hamilton, James, "Long Swings in the Exchange Rate: Are They in the Data and Do Markets Know It?," *American Economic Review*, forthcoming.

Frankel, Jeffrey, and Froot, Kenneth, "Chartists, Fundamentalists, and the Demand for Dollars," in A. Courakis and M. Taylor, eds., *Policy Issues for Interdependent Economies*, London, forthcoming.

______ and ______, "Exchange Rate Forecasting Techniques, Survey Data, and Implications for the Foreign Exchange Market," working paper, International Monetary Fund, March 1990.

Friedman, Milton, "The Case for Flexible Exchange Rates," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Froot, Kenneth, and Frankel, Jeffrey, "Forward Discount Bias: Is It an Exchange Risk Premium?," *Quarterly Journal of Economics*, February 1989, *104*, 139–61.

Krugman, Paul, "Is the Strong Dollar Sustainable?," in *The U.S. Dollar — Recent Developments, Outlook and Policy Options*, Federal Reserve Bank of St. Louis, 1985, 103–33.

Volatility in the Foreign Exchange and Stock Markets: Is It Excessive?

By ROBERT J. HODRICK*

This paper examines recent theoretical and empirical analyses that may help us understand and eventually model the observed movements of foreign exchange rates and stock prices. Whether these prices are rational, efficient, and hence not excessively volatile is, of course, a joint hypothesis. The question in the title can never be answered in the abstract. We must always compare actual volatility to the volatility implied by a benchmark model.

I. What Are Market Fundamentals?

The debate therefore must be about the nature of market fundamentals, which are the economic forces that models with rational behavior predict are the determinants of asset prices. For the stock market, simple rational models imply that real stock prices are the discounted expected values of future real dividends. Hence, market fundamentals are the future dividends that an investor will receive and the current and future discount rates. Notice that this definition says nothing about the stationarity or ergodicity of dividends or dividend growth, the presence or absence of taxes, how market participants learn, or why required rates of return move over time.

Unfortunately, conducting inference about the volatility of asset prices does require explicit assumptions about these issues, especially about the relation of the past to what agents think the future will be like. Robert Flood and I (1990) discuss the importance of correctly identifying market fundamentals in the context of testing for bubbles in asset prices. We stressed that tests for bubbles are interesting specification tests, but must be interpreted with care. This may be telling us only that people expect the future to be different from the past.

Determination of market fundamentals for exchange rates requires general equilibrium considerations since exchange rates play many roles in an economy. Surely, the demand for and supply of money are important for we know what happens to money supplies and exchange rates during hyperinflations. Apparently, as Richard Meese and Andrew Rose discuss in this session, such influences are not center stage currently for movements in the values of major currencies.

Changes in exchange rates also change the relative prices of goods, which affects trade balances and the accumulation of net foreign assets. One market fundamental of exchange rates that is particularly poorly measured is changes of wealth across countries. Meese and Rose use cumulated trade balances to measure net foreign assets, but capital gains are not part of the balance of payments.

Robert Eisner and Paul Piper (1990) examine how U.S. net foreign asset data must be changed if asset values are market-to-market. The Bureau of Economic Analysis reports dollar values of foreign direct investment by multiplying historical book values of U.S. investment overseas by the current exchange rate. No adjustments are made for changes in the local currency values of the investments. Similarly, no changes are made to the dollar values of foreign direct investment in the United States. The Bureau of Economic Analysis (BEA) reports a 1988 U.S. net foreign liability of $532.5 billion, but Eisner and Piper's alternative is $117.1 billion.

Changes in wealth of this magnitude must have important influences on models of exchange rate determination. If Eisner and

---

*Professor of Finance, Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60208, and Research Associate, NBER, Cambridge, MA 02138.

Piper are correct, it is much less surprising that the dollar strengthened in 1989. This is because additional depreciation of the dollar to improve the competitive position of U.S. industries to in turn prevent additional accumulation of debt was not necessary. Such adjustments strengthen the case for rationality of foreign exchange markets and for the modeling of capital flows as described in my 1989b paper.

II. Are Agents Risk Neutral?

Early papers on stock price and exchange rate volatility assumed that economic agents are risk neutral, which implies that expected returns on assets are equal and constant in real terms. Although this greatly simplifies the problem of determining market fundamentals, it does not imply that nominal stock prices or exchange rates follow a random walk, with or without drift. Without risk neutrality, there is no presumption that asset returns are expected to be the same either across assets or over time. Time-varying risk premiums complicate studies of excess asset price volatility because things that change risk premiums on assets are legitimate market fundamentals.

My 1987 book surveys some statistical evidence about the nature of risk premiums in the foreign exchange market. An important message is that we do not have a model of the risk premium that is not rejected by the data. Hence, we have three possible alternative interpretations of the data, each of which may be influencing the statistical analysis.

One possibility is that the market is inefficient. In this case the expected return to speculation is more than sufficient to compensate speculators for bearing risk. Although volume in the foreign exchange market is estimated to be over $130 billion per day, which is prima facia evidence that people are taking risks, turnover per se is unimportant. Open positions must be taken if risk is to be borne, and we have little data on this issue.

A second interpretation is that there is learning in the markets. Karen Lewis (1988; 1989a, b) explores various aspects of learning in the foreign exchange market. The possibility that learning is important received empirical support from Jeffrey Frankel and Kenneth Froot (1987), and the interested reader is referred to their paper in this session for additional insight.

A third possibility is that the data do not satisfy the statistical assumptions that we must employ in order to conduct inference. Nothing in economic theory guarantees that government policies or the choices of economic agents produce time-series that satisfy the assumptions of stationarity and ergodicity.

III. Recent Theoretical Advances

Interesting new theoretical models examine stochastic variation in the conditional variances of traditional market fundamentals as sources of variability in expected returns and asset prices. Andrew Abel (1988) parameterizes an asset pricing model of Robert Lucas (1978) to allow serially correlated dividends and conditional variances of dividends. Abel finds "that the magnitudes of the effect on stock prices of increased dividend volatility is an increasing function of the persistence of the increase in volatility" (p. 385). He assumes that the representative agent maximizes a utility function that is separable across time and has constant relative risk aversion with coefficient α. Many of his results, such as the response of stock prices to a change in dividend volatility, hinge on whether α is greater or less than one.

Alberto Giovannini (1989) extends Abel's model to a monetary economy as in the analysis of Lars Svensson (1985). Otherwise, the specification is similar to Abel's, yet the results are surprisingly different. Giovannini notes that, in this class of models, α is also the reciprocal of the elasticity of intertemporal substitution. Abel finds that increases in dividend variability decrease stock prices if $\alpha<1$ and increase stock prices if $\alpha>1$. Giovannini finds just the opposite. He reconciles the difference in findings by noting that large α's imply low intertemporal substitution.

An increase in uncertainty about future consumption increases precautionary saving when substitutability of consumption across periods is low. In such a situation, if agents could decrease current consumption, they would; but in general equilibrium, current consumption is given by the nonstorable endowment. Hence, for savings to rise in Abel's model, stock prices rise. In Giovannini's model, there can also be a portfolio shift out of stocks and into money. Stocks are less liquid than money and cannot be used to finance consumption in the next period. Hence, when increased uncertainty about the future causes a desired increase in saving to finance consumption in the next period, stock prices fall as agents shift out of stocks and into money.

My 1989a paper extends these results to a two-country model in which the monetary and endowment processes are conditionally heteroscedastic. Hence, exchange rates, interest rates, stock prices and all risk premiums depend on the conditional variances of the exogenous processes. I impose identical tastes across countries, which allows for an analytical solution using the perfectly pooled equilibrium first employed by Lucas (1982). Unfortunately, this solution also imposes purchasing power parity, which we know is grossly violated in the data.

IV. Recent Empirical Analyses of Stock Price Volatility

Recent empirical analyses that address the determinants of stock price variability now recognize that expected returns vary over time and that it is more likely to be true that the price-dividend ratio is a stationary variable than that the level of price or detrended price is stationary.

For example, John Campbell and Robert Shiller (1988a, b) develop a log-linearized vector-autoregressive (VAR) methodology to test asset pricing models. Observable candidate determinants of expected returns on stocks enter the VAR with the logarithm of the price-dividend ratio and dividend growth. Campbell-Shiller test the asset pricing models and reject them all, but they also calculate the variance of the price-dividend ratio implied by various models. Typically, the standard deviations of the implied price-dividend ratios are one-half to two-thirds that of the actual series. A major finding is that variation in the expected return on stocks relative to short-term interest rates will be necessary to reconcile models with the data.

Campbell-Shiller (1989) also conduct a Monte Carlo analysis of their methods. While many of the simulated test statistics are well-behaved, several that rely on nonlinear transformations are not. The results are also sensitive to the presence of near unit roots in the vector autoregression.

Other empirical analysis that is providing interesting and controversial results is the literature on the long-horizon predictability of stock returns. Eugene Fama and Kenneth French (1988a, b; 1989), James Poterba and Lawrence Summers (1988), Andrew Lo and A. Craig MacKinlay (1988) and others document return predictability at horizons from one week to several years.

Fama and French use a regression methodology and find that variables like dividend yields, the yield spread in the term structure of interest rates, and the junk bond default premium have predictive power for returns on portfolios of stocks and bonds.

Poterba-Summers and Lo-MacKinlay both employ a variance-ratio methodology. If returns were independently and identically distributed, the ratio of the variance of returns over k months, when divided by k, to the variance of returns over 12 months, when divided by 12, would be one. For the NYSE value-weighted real returns for the period 1926–85, Poterba-Summers find values at the 48-month horizon of 0.747 (0.232) and at the 96-month horizon of 0.575 (0.394) with Monte Carlo standard errors in parenthesis. For the NYSE equal-weighted real returns, the corresponding values are 0.745 (0.232) and 0.353 (.394).

Notice that none of these test statistics is more than two standard deviations from zero. Poterba-Summers argue that if one insists on applying a .05 probability of a Type I error when testing the null hypothesis that returns are serially uncorrelated, there is often a very high probability of committing a Type II error. Hence, even though most of their estimates fail to fall more than two standard deviations away from the value implied by the null hypothesis of constant expected returns, they reject the null hypothesis. This is because the point estimates are similar to values that are produced when as much as three-fourths of the variance in returns is due to highly serially correlated transitory components in stock prices, and because of the consistency of the evidence across a variety of data sets.

This literature is controversial as evidenced by the papers of Myung Kim et al. (1989), Matthew Richardson (1988), and James Stock and Richardson (1989). Kim et al. argue primarily that the results of autocorrelation based tests are very sensitive to inclusion of the depression and that there is no evidence of negative serial correlation in returns when it is excluded. Richardson and Stock-Richardson argue, among other things, that the small sample properties of these test statistics are poor.

My paper with Lars Hansen (1989) examines alternative ways of doing inference and measurement about long-horizon forecasting issues using only one-step ahead information. We found that long-horizon test statistics may have poor small sample properties, but the most interesting results are from one-step-ahead vector autoregressions of returns and dividend yields. The dynamics of the VAR imply long horizon R^2s of returns measured as one minus the ratio of the innovation variance of the sum of returns over k months to the total variance of the sum of returns over k months. For the period 1948–87, the one-step-ahead R^2 is between 3 and 11 percent depending on the order of the VAR, but the R^2s always rise to over 40 percent for the sum over 48 months ahead, regardless of the order of the VAR. Monte Carlo simulations indicate that such findings are very unlikely to be produced by chance.

These results are related to findings of Campbell (1989) who develops measures of the persistence of expected returns from similar VAR models. Campbell asks how much must the current stock price fall when the expected return rises in order that the dynamics of expected returns are consistent with the data. With monthly data for 1952 to 1988, he finds a point estimate of a 5.8 percent fall for a 1 percent innovation in the current expected return.

V. Conclusions

One message of this paper is that movements in conditional variances of market fundamentals are legitimate market fundamentals that imply both movements in expected returns and movement in conditional variances of asset prices. Richard Baillie and Tim Bollerslev (1989) document movements in conditional variances for foreign exchange rates and French et al. (1987) provide analysis for the stock market. Integrating these analyses into economic models is a challenging area for future research.

I think full explanations of observed volatilities in these markets may require models with differential information across agents and learning about government policies and the nature of technology. I say this because many economists seem to think there is no rational explanation for the October 1987 worldwide stock market crash, and my own explanation relies on these features as in the following scenario.

Substantial bad news about the U.S. trade balance was accumulating. This increased the likelihood of a trade war that would be bad for the profitability of corporations in all countries and increased speculation that foreign investors would dump dollar assets because the dollar would have to depreciate to correct the trade balance. The market needed answers to questions like could and would the Federal Reserve defend the dollar, would such attempts increase real interest rates and create a recession, how badly would a recession aggravate the budget deficit, and what would be the response of foreign governments to these events. Differences of opinion across market participants about the answers to these questions caused trading volume to increase by much more than the market thought could be done with existing technology. The combination of the increase in bad news and the increase in volume and volatility forced stock prices to fall, partly in response to the bad news and partly to produce higher expected returns. Fortunately, no trade was started, no major recession ensued, and reasonable government policies were followed around the world.

REFERENCES

Abel, Andrew B., "Stock Prices Under Time-Varying Dividend Risk: An Exact Solution in an Infinite-Horizon General Equilibrium Model," *Journal of Monetary Economics*, November 1988, *22*, 375–93.

Baillie, Richard T. and Bollerslev, Tim, "The Message in Daily Exchange Rates: A Conditional Variance Tale," *Journal of Business and Economics*, July 1989, *7*, 297–305.

Campbell, John Y., "Measuring the Persistence of Expected Returns," manuscript, London School of Economics, 1989.

______ **and Shiller, Robert J.,** (1988a) "Stock Prices, Earnings and Expected Dividends," *Journal of Finance*, July 1988, *43*, 661–76.

______ **and** ______, (1988b) "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, Fall 1988, *1*, 195–228.

______ **and** ______, "The Dividend Ratio Model and Small Sample Bias," *Economic Letters*, 1989, *29*, 325–31.

Eisner, Robert and Piper, Paul J., "The World's Greatest Debtor Nation?," *Review of Economics and Finance*, forthcoming 1990.

Fama, Eugene F. and French, Kenneth R., (1988a) "Permanent and Temporary Components of Stock Prices," *Journal of Political Economy*, April 1988, *96*, 246–73.

______ **and** ______, (1988b) "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, October 1988, *22*, 3–26.

______ **and** ______, "Business Conditions and Expected Returns on Stocks and Bonds," CRSP Working Paper No. 220, University of Chicago, 1989.

Flood, Robert P. and Hodrick, Robert J., "Testable Implications of Indeterminacies in Models with Rational Expectations," *Journal of Economic Perspectives*, May 1990.

Frankel, Jeffrey A. and Froot, Kenneth A., "Using Survey Data to Test Standard Propositions Regarding Exchange Rate Expectations," *American Economic Review*, March 1987, *77*, 133–53.

French, Kenneth R., Schwert, G. William and Stambaugh, Robert F., "Expected Stock Returns and Volatility," *Journal of Financial Economics*, September 1987, *19*, 3–30.

Giovannini, Alberto, "Uncertainty and Liquidity," *Journal of Monetary Economics*, March 1989, *23*, 239–58.

Hodrick, Robert J., *The Empirical Evidence on the Efficiency of Forward and Futures Foreign Exchange Markets*, Chur: Harwood Academic Publishers, 1987.

______, (1989a) "Risk, Uncertainty, and Exchange Rates," *Journal of Monetary Economics*, May 1989, *23*, 433–59.

______, (1989b) "U.S. International Capital Flows: Perspectives from Rational Maximizing Models," in Karl Brunner and Allan H. Meltzer, eds., *Carnegie-Rochester Conference Series on Public Policy: International Debt, Federal Reserve Operations, and Other Essays*, August 1989, *30*, 231–88.

______ **and Hansen, Lars Peter,** "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement," manuscript, Northwestern University, 1989.

Kim, Myung Jig, Nelson, Charles R. and Startz, Richard, "Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence," manuscript, University of Washington, Seattle, 1989.

Lewis, Karen K., "The Persistence of the 'Peso Problem' When Policy Is Noisy," *Journal of International Money and Finance*, March 1988, *7*, 1–11.

______, (1989a) "Can Learning Affect Exchange Rate Behavior? The Case of the Dollar in the Early 1980's," *Journal of Monetary Economics*, January 1989, *23*, 79–100.

______, (1989b) "Changing Beliefs and Systematic Rational Forecast Errors with Evidence from Foreign Exchange," *American Economic Review*, September 1989, *79*, 621–36.

Lo, Andrew W. and MacKinlay, A. Craig, "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, Spring 1988, *1*, 41–66.

Lucas, Robert E., Jr., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, *46*, 1429–45.

______, "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, May 1982, *10*, 335–60.

Poterba, James M. and Summers, Lawrence H, "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, October 1988, *22*, 27–59.

Richardson, Matthew, "Temporary Components of Stock Prices: A Skeptic's View," manuscript, Stanford University, 1988.

Stock, James H. and Richardson, Matthew, "Drawing Inference from Statistics Based on Multi-Year Asset Returns," manuscript, Harvard University, 1989.

Svensson, Lars E. O.,, "Money and Asset Prices in a Cash-in-Advance Economy," *Journal of Political Economy*, October 1985, *93*, 919–44.

Nonlinear, Nonparametric, Nonessential Exchange Rate Estimation

By RICHARD A. MEESE AND ANDREW K. ROSE*

A wide variety of empirical exchange rate models have been estimated over the years. But, despite the considerable energies that have been devoted to this work, the economics profession has remarkably little to show for itself. There is little evidence that conclusively links the bilateral exchange rates of typical OECD countries to "fundamental" macroeconomic determinants of exchange rates, such as money, output, relative prices, or interest differentials. Coefficient estimates are notoriously unstable and frequently "mis-signed" (compared with theoretical predictions); exchange rate equations do not fit particularly well, and forecast no better than the simplest naive alternatives.

Recently, a new class of exchange rate models was introduced by Paul Krugman (1988). These models provide a potential reason for the poor performance of traditional exchange rate models, because they are nonlinear. If the exchange rate actually depends in a nonlinear way on exogenous macroeconomic fundamentals, linear exchange rate models may work poorly, even though the exchange rate is closely linked to fundamentals. In this paper we provide a brief sketch of some of this work, as well as some preliminary evidence on the actual performance of these nonlinear models.

In our empirical analysis, we use a nonparametric estimator that can handle a wide variety of nonlinear phenomena. We examine fixed exchange rate regimes, where nonlinearities should be quite easy to detect. However, we do not find strong empirical support for the hypothesis that the incorporation of nonlinear effects significantly improves models of exchange rate determination.

In Section I, we briefly review the theoretical literature on nonlinear "target zone" exchange rate models, linking this work to the tests for "intrinsic" bubbles (we draw heavily on recent papers by Kenneth Froot and Maurice Obstfeld, 1989a, b). Our methodology and data are discussed in Section II; Section III contains new empirical tests for nonlinearities in exchange rate models.

I. Target Zones and Intrinsic Nonlinearities

Froot and Obstfeld (1989a) provide a good treatment of the target zone literature in the context of the static flexible price monetary exchange rate model. This assumes identical log-linear money demand functions at home and abroad, and exogenous, random deviations from both purchasing power parity (PPP) and uncovered interest parity.

Define the composite driving "fundamental" process as

$$(1)\quad x_t = a_1[y_t^* - y_t] + [m_t - m_t^*] + u_t,$$

where a_1 is the income elasticity of money demand, y denotes domestic real income, m denotes nominal domestic money, and u_t is a composite white noise disturbance (the aggregate of the four shocks to domestic and foreign money demand, the PPP deviation, and the deviation from uncovered interest parity). An asterisk denotes a foreign value. The equilibrium exchange rate can be expressed as

$$(2)\qquad e_t = x_t + a_2 E[de_t/dt|I_t]$$

where e_t is the domestic price of foreign currency, a_2 is the interest semi-elasticity of money demand, $E(\cdot|I_t)$ denotes the expectations operator, conditional on information

---

*350 Barrows Hall, University of California, Berkeley CA 94720. For assistance in obtaining data, we thank the Center for Research and Management, the Institute for Business and Economic Research, and the International Monetary Fund. For comments, we thank Joe Gagnon and Jonathan Ostry. Our data and programs are available upon request.

I_t, and de_t/dt denotes the time derivative of the exchange rate. This relation is written in continuous time in order to exploit new solution techniques.

Suppose that during a regime of cleanly floating exchange rates, the fundamentals evolve as a continuous time random walk with a drift of rate n:

$$x_t = n\,dt + a_3\,dz_t, \tag{3}$$

where dz_t is a standard Wiener process (more complicated x processes can also be considered). When (3) is expected to hold permanently, the solution for e_t is given by the familiar "present value" formula:

$$e_t^{PV} = x_t + a_2 n. \tag{4}$$

However, if policymakers adjust fundamentals to prevent the exchange rate from leaving the "reflecting bands" of a prespecified "target zone" (e^L, e^U), then (3) (and hence (4)) is inappropriate. A solution for the exchange rate can be obtained by applying the techniques of regulated Brownian motion (see, for example, J. Harrison, 1985; Krugman, 1988; and Froot-Obstfeld, 1989a).

Here we paraphrase the two-step solution algorithm of Froot-Obstfeld (1989a). The first step is to characterize the family of functions $e = G(x)$ that satisfy (2) when the driving process for x follows (3). The second step is to identify a particular function in the family $G(x)$ that accommodates the target zone boundary conditions. Without further detail we assert that the target zone solution to (2) is

$$e_t^{TZ} = x_t + a_2 n + A_1 \exp(b_1 x_t) + A_2 \exp(b_2 x_t), \tag{5}$$

where b_1 and b_2, as well as the constants of integration A_1 and A_2 are functions of e^L, e^U, and the parameters of (1)–(3). The two nonlinear terms enter because of the target zone policy and are the sole difference between (4) and (5).

It is widely believed that the monetary authorities of Western countries have attempted to limit exchange rate variability since the Louvre Agreement; Krugman (1989) argues that fixed exchange rate regimes (such as the EMS) can also be interpreted as target zones with very narrow bands. Thus, nonlinear exchange rate effects may be of great empirical importance. They also provide a potential explanation for the failure of linear exchange rate models; if the authorities were expected to defend target zones, the omission of nonlinear effects constitutes model misspecification.

Nonlinear exchange rate effects need not arise only because of target zones; (nonlinear) "bubbles" can also give rise to the nonlinear expressions such as those in (5). When agents are certain that a free float will continue forever, the linear solution, (4), is appropriate. In this case, the nonlinear terms in (5) can be interpreted as bubbles that are "intrinsic" in that they depend only on the fundamentals that are an integral part of the linear solution (4).

This type of bubble has a number of desirable features that "extrinsic" bubbles lack. Both types of bubbles are driven by self-fulfilling expectations, but extrinsic bubbles depend on extraneous variables not related to fundamentals. Froot and Obstfeld (1989b) contrast the two types of bubbles, and demonstrate that intrinsic bubbles have greater potential to explain some of the shortcomings of current asset pricing models like (2).

Extant empirical work on bubbles in currency markets has focused mainly on tests for extrinsic bubbles; see Meese (1986) among others. These tests can have low power against plausible alternative models, and test rejections can be interpreted as evidence of other forms of misspecification of the exchange rate model. For example, in the present context, nonlinearities in the empirical exchange rate model we consider may not be manifestations of either target zones or intrinsic bubbles, since our tests for nonlinearities rely on the assumption that (2) and (3) are correctly specified. Our earlier paper (1989) provides further analysis of different sources of nonlinear phenomena, including functional form misspecification and time deformation.

II. Nonparametric Regression Methodology and Data

We attempt to uncover nonlinear effects using the recently developed nonparametric "Locally Weighted Regression" (LWR) technique. The LWR is a technique for estimating regression surfaces in a moving average manner; an extremely wide range of functions can be detected with the technique. The LWR includes the nonparametric "nearest-neighbor" technique as a special case. William Cleveland and Susan Devlin (1988) provide a variety of examples and related theoretical results.

Suppose that the regression model is given by

$$(6) \qquad e_t = f(x_t) + w_t \qquad t = 1, \ldots, T$$

where x_t is a vector of weakly exogenous variables, $f(\cdot)$ is a smooth function, and w_t is an i.i.d. disturbance, normally distributed with mean zero and finite variance. The object of interest is an empirical estimate of $f(x')$.

The LWR uses a "window" of the k observations closest to x', where proximity is assessed using the Euclidean distance between x' and x_t, denoted $D(x', x_t)$. The LWR is simply weighted least squares of e on x for the relevant k observations, where the weights are given by $W = V[D(x', x_t)/D(x', x_k)]$. The distance from x' to the kth closest x_t is $D(x', x_k)$. We follow the suggestion of Cleveland-Devlin, and use the steep "tricube" function for $V(\cdot)$: $T(v) = (1 - v^3)^3$ if $v < 1$, and 0 otherwise. Since the LWR estimator of $f(\cdot)$ is linear in e, the statistical properties of the estimator can be worked out with standard techniques. The analysis is somewhat complicated insofar as the error sum of squares does not have an exact *chi*-squared distribution, and must be approximated by a method of moments technique.

In our earlier paper, we tested for nonlinearities in a variety of structural models of exchange rate determination over the post-Bretton Woods regime of generalized floating. We used three different statistical techniques for four bilateral dollar exchange rates but could not find strong evidence of economically or statistically important nonlinearities. This finding is consistent with the absence of target zone effects, nonlinear intrinsic bubbles and other specification errors. However, target zones were not perceived to be in effect for much of the sample. Following the analysis of Krugman (1989) that identifies fixed exchange rate regimes as target zones with narrow bands, we now test for nonlinearities in fixed exchange rate regimes. We do this in order to give the data the best possible chance of revealing target zone nonlinearities.

Our main data set is drawn from the Bretton Woods regime of fixed but adjustable pegs, which prevailed after World War II until the early 1970s. We use continuous samples of data from the 1960s, excluding periods that include major exchange rate realignments from our sample (for instance, our British sample ends before the sterling devaluation of 1967). The data are monthly and seasonally adjusted. We estimate flexible-price monetary models, using industrial production as an output proxy, and $M1$ for the money supply. This allows us to model nominal bilateral exchange rates between the United States and Canada, France, Germany, Italy, Japan, Netherlands, Sweden, Switzerland, and the United Kingdom.

To circumvent the obvious problem of money supply endogeneity in a fixed rate regime, we estimate equation (4) by OLS and LWR while incorporating relative monies (with unit coefficients) into the dependent variable. Also, Robert Aliber (1978) provides evidence on the validity of the "Fisher-Open" hypothesis (that links expected and actual rates of depreciation to the interest differential) during the Bretton-Woods regime.

We also use two other data sets. These data sets are of higher frequency and are therefore univariate. One data set is drawn from the gold standard regime that prevailed before World War I, and consists of weekly nominal bilateral exchange rates between the United States and three of its major trading partners: France; Germany; and the United Kingdom. The data are the midpoint of the weekly high and low values (using only the high or low points does not dramatically change results). The data are available from 1899 through 1909.

Our last data set consists of daily exchange rate quotations from the EMS. Our data are midpoints of *DM/Ff* (German/French) and *DM/Fl* (German/Dutch) spot rates, based on daily London noon quotations. The regimes are again chosen to avoid EMS realignments.

III. Empirical Analysis of Fixed Exchange Rates

We compare linear parametric estimates of (4) with nonparametric nonlinear estimates of (5) using the *F*-test suggested by Cleveland and Devlin.[1] The *F*-tests compare the fits generated by the two different estimation techniques; the null hypothesis is that (linear) OLS estimation provides a fit that is not significantly worse than LWR (that can accommodate nonlinearities). A high *F*-test is thus inconsistent with the hypothesis that there are no nonlinearities; that is, indicates the presence of significant nonlinearities.

Throughout, all variables are expressed in first-differences of natural logarithms. We report results for a LWR window size of 80 percent of the sample, and our conclusions appear robust to this choice. Also, our results appear to be insensitive to exact choice of sample endpoints.

Table 1 contains *F*-tests that compare linear and nonlinear estimates of the static flexible-price monetary model. Degrees of freedom are recorded parenthetically in the tables. There are no indications of important nonlinearities at conventional levels of statistical significance. These negative results confirm those of our earlier paper, as well as those of Francis Diebold and James Nason (1989), who find that LWR does not lead to any improvement in out of sample predictive power in exchange rates vis-à-vis the random walk.

TABLE 1—*F*-TESTS FOR NONLINEARITIES IN BRETTON WOODS[a]

Country	Sample	*F*	DF
Canada	1962:8–69:9	0.20	(3,88)
France	1960:4–69:4	1.09	(6,108)
Germany	1961:6–69:6	1.93	(4,99)
Italy	1963:4–69:4	0.53	(4,75)
Japan	1960:4–69:9	0.71	(3,116)
Netherlands	1961:6–69:9	1.26	(3,102)
Switzerland	1960:4–69:9	2.02	(3,116)
Sweden	1963:4–69:9	0.98	(4,79)
United Kingdom	1963:3–67:7	0.35	(3,54)

[a] Monthly data.

TABLE 2—UNIVARIATE *F*-TESTS FOR NONLINEARITIES IN GOLD STANDARD[a]

	U.K./U.S.	France/U.S.	Germany/U.S.
F-test	1.59	3.05	1.89
DF	(2,568)	(5,562)	(5,562)

[a] Weekly data, 1899–1909.

TABLE 3—UNIVARIATE *F*-TESTS FOR NONLINEARITIES IN EMS[a]

Sample Period	—*Dm/Fl*—	—*DM/Ff*—
1979:3:14–79:9:23	1.13(6,113)	0.78(6,117)
1979:9:28–79:11:29	1.06(9,29)	0.98(7,29)
1979:12:3–81:3:21	5.11(7,278)	2.76(6,279)
1981:3:26–81:10:4	1.68(7,108)	0.45(6,79)
1981:10:9–82:2:21	0.51(6,79)	1.34(5,75)
1982:2:26–82:6:13	1.43(6,62)	1.55(6,47)
1982:6:18–83:3:20	2.13(2,161)	2.26(4,140)
1983:3:25–85:7:20	2.19(6,549)	4.20(5,549)
1985:7:25–86:4:6	1.43(2,166)	3.58(3,162)
1986:4:11–86:8:3	0.03(6,73)	1.38(6,72)
1986:8:8–87:1:11	0.43(5,100)	1.68(8,98)
1987:1:16–89:3:1	0.90(5,503)	2.39(5,508)

[a] Daily data.

In Table 2, we report univariate tests for exchange rate nonlinearities in 1899–1909 (the pre-World War I gold standard period). More precisely, the *F*-tests compare the fits of first-order autoregressions estimated with OLS and LWR. One of the three test statistics (for France) indicates important nonlinearities.

We report in Table 3 a series of univariate tests for nonlinearities in the first decade of EMS operations. These *F*-tests compare the fit of second-order linear autoregressions with second-order autoregressions estimated with LWR. The results do not strongly indicate the presence of significant nonlinearities; of the 22 test statistics reported, 5 are inconsistent with the null hypothesis at the .05 level (one of these is significant at the .01 level). Most of the rejections stem from the franc/mark rate.

---

[1] Parenthetically, an alternative to our nonparametric approach would be to estimate (5) and (3) simultaneously in a parametric but nonlinear fashion; we defer this topic to further research.

All of these results seem insensitive to exact choice of sample; reducing the sample by eliminating data closer to exchange rate realignments does not change any results. However, there are occasional indications of nonnormal or autocorrelated residuals. In order to assess the importance of these phenomena, and to evaluate the statistical power of the LWR F-test to detect nonlinearities, we performed a set of simulation experiments with pseudo exchange rate data. These data were generated using the solution (5) and a grid of money demand parameters. The model was calibrated so that the linear terms in the exchange rate solution, $x_t + a_2 n$, accounted for 5 percent of the variation in e_t, a result consistent with in-sample estimation of equation (4) in first differences.

In all cases the value of the nonlinear F-statistic was highly significant in a regression of e_t on x_t and a constant. The average value of the pseudo data F-statistics is also considerably higher than any of the F-statistics we report in Table 1 based on Bretton Woods data. A puzzling aspect of our simulation experiments concerns the univariate autoregressive (AR) representation of the exchange rate process. When an AR(1) model is estimated by LWR with the pseudo data, evidence of nonlinearity is present in only about one-fifth of the experiments. This leads us to doubt the validity of our autoregressive tests for target zones using gold standard data and EMS data.

IV. Conclusion

In this paper, we have looked for nonlinear exchange rates effects in three different regimes of fixed exchange rates. While autoregressive tests for nonlinearities do not appear to have good power characteristics, more powerful tests also do not indicate statistically important nonlinearities in the Bretton Woods fixed exchange rate regime. Further research is clearly warranted. Nevertheless, with negative results in hand, it may be reasonable to suspect the empirical importance of stochastic regime shifts, at least in the context of simple monetary exchange rate models.

REFERENCES

Aliber, Robert, *Exchange Risk and Corporate International Finance*, London: Macmillan, 1978.

Cleveland, William and Devlin, Susan, "Locally Weighted Regression," *Journal of the American Statistical Association*, September 1988, *83*, 596–610.

Diebold, Francis and Nason, James, "Non-Parametric Exchange Rate Prediction?," mimeo., FRB, 1989.

Flood, Robert and Garber, Peter, "Market Fundamentals versus Price-Level Bubbles," *Journal of Political Economy*, August 1980, *88*, 745–70.

Froot, Kenneth and Obstfeld, Maurice, (1989a) "Exchange Rate Dynamics under Stochastic Regime Shifts: A Unified Approach," mimeo., MIT, 1989.

______ **and** ______, (1989b) "Intrinsic Bubbles: The Case of Stock Prices," mimeo., MIT, 1989.

Harrison, J., *Brownian Motion and Stochastic Flow Systems*, New York: Wiley, 1985.

Krugman, Paul, "Target Zones and Exchange Rate Dynamics," NBER Working Paper No. 2481, 1988.

______, "Target Zones with Limited Reserves," mimeo, MIT, 1989.

Meese, Richard, "Testing for Bubbles in Exchange Markets," *Journal of Political Economy*, April 1986, *94*, 345–73.

______ **and Rose, Andrew,** "An Empirical Assessment of Non-Linearities in Models of Exchange Rate Determination," IFDP no. 367, 1989.

INTERACTION BETWEEN ENVIRONMENTAL AND AGRICULTURAL POLICIES: OPPORTUNITIES FOR COORDINATION AND LIMITATIONS FOR EVALUATION†

Interactions Between Agricultural and Environmental Policies: A Conceptual Framework

By RICHARD E. JUST AND JOHN M. ANTLE*

Agricultural policy has a well-documented impact on farmers' production decisions, and those decisions (land, water, and agricultural chemical use) may in turn affect the environment. There is a public perception that existing agricultural policies are linked to agricultural pollution, as exemplified by the conclusions of the recent National Research Council report, *Alternative Agriculture* (1989). There is only limited theoretical or empirical research addressing these linkages, however.

This paper develops a conceptual framework that can be used to analyze the interactions between agricultural and environmental policies and pollution. This framework integrates physical and economic models at a disaggregate level necessary to capture the heterogeneity of the physical environment and the economic behavior of farmers. Following E. Hochman and D. Zilberman (1978), the disaggregate units can be statistically aggregated to a level useful for policy analysis. Agricultural and environmental policies can be categorized according to their effects on the intensive and extensive margins. Combining the different policy instruments with the microeconomic heterogeneity, we find that existing agricultural and environmental policies can have either positive or negative effects of nonpoint source pollution; to infer an aggregate effect requires data that do not currently exist. Our framework provides a basis for empirical investigations of these effects, given adequate data. Moreover, we find that agricultural policy instruments can be used to mitigate pollution if used appropriately.

I. A Disaggregated Model

Consider a region defined in relation to an environmentally meaningful geographical unit, such as a watershed or aquifer. The jth acre in the region has a set of environmental characteristics ω_j, that affect both its agricultural productivity and the production of pollution. A variety of complex physical models are being developed to measure pollution caused by agricultural production, such as surface and ground water contamination. The stylized physical model here is represented by the function $z_j = z(x_j, \omega_j)$, where x_j is the level of input use on the jth acre and z_j is pollution generated by production on the jth acre.

The economic model is based on the optimal allocation of land and other inputs in production as functions of prices, policies, and the environmental characteristics of the land managed by the farmer. To focus on the role of land quality, all farmers are assumed to be risk neutral and to produce with identical technology. Farms are differentiated only by the environmental characteristics of their land (a more general model would include a vector of farm-specific characteristics). In the production period, the ith farmer manages n^i acres with environmental characteristics $\omega^i = (\omega_1^i, \omega_2^i, \ldots)$. Define the indicator function δ_j^i such that $\delta_j^i = 1$ if acre j is in production and 0 otherwise, and let $\delta_i = \{\delta_j^i\}$. The vector of attributes of land in production on farm i is then $\omega(\delta_i) = (\omega_1^i \delta_1^i, \omega_2^i \delta_2^i, \ldots)$ and total acreage in production on the ith farm is $\Sigma_j \delta_j^i$.

---

†*Discussants*: W. Michael Hanemann, University of California-Berkeley; Katherine Reichelderfer, U.S. Department of Agriculture; Nancy E. Bockstael, University of Maryland.

*Professor of Agricultural and Resource Economics, University of Maryland, College Park, MD 20742, and Senior Staff Economist, President's Council of Economic Advisors, Washington, D.C. 20500, respectively.

All farms in the region face the same vectors p and q of prices and policy parameters. Define x_j^i as the input allocation of farmer i to acre j and x^i as the vector of x_j^i. The ith farmer's decision problem can then be cast as

$$\max_{x^i, \delta^i} \pi\left[x^i, \omega(\delta^i) \,|\, p, \psi, \omega^i\right] \tag{1}$$

where π is the farmer's objective function embedding the production technology and ψ is a vector of policy parameters.

The solution to this maximization problem generates the demand functions $x_j^i = x(p, \psi, \omega_j^i)$ and $\delta_j^i = \delta(p, \psi, \omega_j^i)$. The environmental characteristic of each unit of farm land in the region is fixed at a point in time and can be viewed as being distributed across the acres in the region with a distribution defined by the parameter vector θ. The distribution of environmental attributes induces a joint distribution for input use x^i and land use δ^i in the region. The environmental attributes of the land in production, $\omega(\delta^i)$, are determined by land use decisions. Yield and pollution are functions of input use and the environmental attributes of the land in production. Thus, farmers' production decisions generate a joint distribution of output, input, environmental attributes, and pollution in the region.

II. Modeling the Joint Distribution of Inputs, Environmental Characteristics, and Pollution

Policy may impose restrictions on the distribution of x and ω that must be taken into account in analysis and estimation. When land-use restrictions limit the range of environmental attributes available for production, the distribution will be truncated in the ω dimension. The distribution of x and ω also may be censored, as when there is a positive probability that input use occurs at zero. This occurs, for example, when pesticide-use decisions may be zero with a positive probability. Similarly, the distribution may be censored at a positive limit, as when policy limits water or chemical use. When truncation or censoring are not important, it is possible to greatly simplify the modeling by assuming a common continuous distribution such as a joint lognormal distribution.

The implementation of an environmental policy such as a restriction on input use per acre may alter farmer behavior, and induce other changes in the distribution $\phi(x, \omega)$ in addition to its censoring. According to the optimization problem defined above, farmers jointly choose which acres to place in production and the inputs used on those acres as functions of environmental attributes of the land. If the jth acre is profitable at $x_j > x_0$, but not at $x_j = x_0$, then that acre would not be put in production under a restriction at x_0. Hence, the resulting joint distribution $\phi(x, \omega)$ for acres in production would also be different. Suppose, for example, that ω and x are positively correlated, and all acres with $\omega > \omega^*$ were unprofitable with $x = x_0$. The policy would thus result in the truncation of the distribution at ω^* for values of $x > x_0$. As a result, the distribution would shift toward the origin in the ω dimension and would also be truncated at x_0. It can be concluded, therefore, that policies can have complex effects on the joint distribution of x and ω. Similar conclusions can be drawn for the joint distribution of inputs and pollution.

III. Policy Interaction at the Extensive and Intensive Margins

In another paper (1990), we apply the general model above to the analysis of simple stylized models indicating some of the general effects of major agricultural and resource policy instruments, while focusing on extensive and intensive margin effects independently. In reality, agricultural policies are composed of a complex set of instruments that interact with one another in determining decisions on the extensive and intensive margins simultaneously. Indeed, the producer's choice problem is defined in Section I as the joint determination of land use and input use. Most policies that affect the economic decisions of farmers thus affect decisions at both the extensive and intensive margins.

In this context, it is interesting to note that major current agricultural policies are structured so as to allow control over both margins. This is important from a resource or environmental policy perspective for the following reason. Agricultural price supports have been criticized because they tend to encourage more intensive farming practices on the acreage in production. Indeed, the analysis of input use holding land use fixed verifies that price supports will generally increase input use and pollution. The analysis of land-use decisions holding input intensity fixed, however, shows that agricultural policies do not necessarily lead to higher pollution. This can be true in those cases in which agricultural productivity and the environmental attributes associated with pollution are negatively related. The point of this section is to demonstrate that current agricultural policies are structured so that undesirable environmental effects can be mitigated if policies are appropriately designed and administered.

Let us now assume that acres diverted from production receive a payment of $g per acre, and there is a diversion requirement of λ percent, or of $n\lambda$ acres. If an acre is put into production, input use is x_j^*. The solution to the land-use problem is obtained by selecting into production those acres that are more profitable than g, while meeting or exceeding the diversion requirement. We assume there is a monotonic relationship between input use and profitability as environmental attributes are varied. Ordering all acres from least to most profitable is equivalent to ordering them from low to high values of x_j^*. The farmer will divert the jth acre if profit π_j does not exceed g, or if it is the least profitable acre with $\pi_j > g$ that must be diverted to meet the requirement. The acreage diversion thus determines a minimum level of input use $\bar{x}$, formally defined as

$$\bar{x}(g,\lambda) \equiv \min\Big\{ x_j^* | \pi_j \le g \quad \text{or } \textstyle\sum_j (1-\delta_j) \ge n\lambda \Big\}.$$

Note that if the marginal acre meeting the diversion requirement is more profitable than the diversion payment, the farmer will stop diverting land at that point. But, if the marginal acre is not profitable, the farmer will exceed the diversion requirement up to the break-even point, or a payment or diversion limitation. Thus, the diversion requirement defines a lower bound on $\bar{x}$, but as g increases holding λ constant, $\bar{x}$ may increase as more profitable land is diverted.

Truncating input use through land diversion will have an impact on the joint distribution of input and pollution, and thus on expected pollution. The effect on expected pollution depends on the correlation between input and pollution. If x and z are positively correlated, removing land with $x < \bar{x}$ also removes land associated with low pollution levels. If x and z are negatively correlated, the opposite tends to be true. Formally, for a binding diversion requirement λ,

$$(2)\quad E(z) = (1-\lambda)^{-1} \times \int_0^\infty \int_{\bar{x}}^\infty z f(x,z)\, dx\, dz \equiv Z$$

where $f(x,z)$ is the joint distribution of input and pollution and

$$\lambda = \int_0^\infty \int_0^{\bar{x}} f(x,z)\, dx\, dz.$$

It follows that

$$d\lambda / d\bar{x} = \int_0^\infty f(\bar{x}, z)\, dz, \quad \text{and}$$

$$\partial Z / \partial y = (1-\lambda)^{-1} [Z - E(z|x=\bar{x})].$$

If z and x are positively correlated, $Z = E(z|x > \bar{x}) > E(z|x = \bar{x})$, and therefore, $\partial Z/\partial\lambda > 0$. Conversely, $\partial Z/\partial\lambda < 0$ if z and x are negatively correlated. Observing that the effect on $\bar{x}$ of an increase in g is the same as an increase in λ, it follows that if λ is not binding then changes in g can also have either positive or negative effects on expected pollution depending on the correlation between x and z.

A price support and land diversion policy is often combined with various forms of environmental policies. For example, pesticide restrictions are often imposed in the form of a uniform standard. To determine the effects of a change in an input restriction x_0 on expected pollution, differentiate expected pollution with respect to x_0:

$$\partial E(z)/\partial x_0 = \int_0^{\infty} \partial z(x_0,\omega)/\partial x_0 \phi(x,\omega)\, dx\, d\omega.$$

It follows that the effect of x_0 on expected pollution is determined by the magnitude and sign of the effect of x on z holding ω constant.

The combined effects of a price support/acreage diversion policy and an input restriction on pollution can now be analyzed. The land diversion induces farmers to remove land with $x < \bar{x}$ from production, and thus to truncate the distribution at $\bar{x}$. The restriction on input use greater than x_0 censors the distribution at x_0, so the mass of the distribution accumulates at x_0. The combined effects of the policies are thus to concentrate the mass of the distribution in the interval $[\bar{x}, x_0]$. If x and z are positively correlated, removing the lower end of the domain of x increases the mean level of pollution and removing the upper end of the domain reduces mean pollution. The opposite occurs if x and z are negatively correlated. These two types of policy therefore work in opposite directions. It can be concluded that if x and z are positively correlated, the preferred policy is to reduce pollution (and production), whereas if x and z are negatively correlated, the preferred policy is an acreage diversion. Conversely, if x and z are positively (negatively) correlated, a diversion (standard) may increase (not reduce) pollution.

The preceding analysis applies to farms that participate in programs. A complete analysis also needs to address the program participation decision. From equation (1), farmer i on acre j earns profit $\pi[p, \varphi, \omega_j^i]$ given prices p, policy parameters φ, and attributes ω_j^i. A farmer that chooses to comply with program rules faces policy parameter vector φ_c, and those who do not, face φ_n. Expected returns on the ith farm are

$$E\left[\pi|\varphi_k, \theta^i\right] = \int \pi\left[p, \varphi_k, \omega\right] \phi(\omega|\phi^i)\, d\omega, \qquad k = c, n.$$

We can define a compliance frontier as those combinations of φ and θ giving $E[\pi|\varphi_n, \theta^i] = E[\pi|\varphi_c, \theta^i]$. Given policy parameters, for example, it is possible to examine the relationship among elements of θ along the frontier. To obtain an expression for aggregate variables, define the distribution of environmental parameters across farms as $f(\theta|\varphi_k)$, so that the proportion of complying and noncomplying farms is $p_k = \int f(\theta|\varphi_k)\, d\theta$, $k = c, n$. Now expected pollution, for example, can be expressed as $E(z) = p_n \int\int z f(x, z)\, dx\, dz + p_c Z$, where Z is defined as in (2).

IV. A Simple Model Combining Extensive and Intensive Margin Effects of Policy

The results of the previous section can be illustrated using a simple example. Let A_i be the total acreage of farm i and let y_i be the average yield on farm i. Suppose yield follows $y_i = a_i(1 - e^{-x_i}) + b_i$ where x_i is the quantity of the polluting (environmentally degradating or resource depleting) input used, and a_i and b_i are constants associated with farm i. Where p_m is the market price for output, v is the price of the input, and c_i is the cost per acre of any other inputs embodied in a_i and b_i, profit on farm i under noncompliance with a voluntary program is

$$\pi_i = A_i\left[p_m a_i(1 - e^{-x_i}) + p_m b_i - v x_i - c_i\right].$$

The condition for profit maximization in this case is

$$\partial \pi_i / \partial x_i = A_i\left(p_m a_i e^{-x_i} - v\right) = 0$$

which implies an optimal input per acre under noncompliance of $x_i^n = \ln(p_m a_i/v)$ if $\pi_i^1 \geq \pi_i^0$, and 0 otherwise, where π_i^1 is the optimal profit when optimal input use is positive,

$$\pi_i^1 = A_i\big[p_m(a_i + b_i) - v - v\ln(p_m a_i/v) - c_i\big],$$

and π_i^0 is the optimal profit when optimal input is zero, $\pi_i^0 = A_i(p_m b_i - c_i)$. Here, one finds

$$\pi_i^1 - \pi_i^0 = A_i\big[p_m a_i - v - v\ln(p_m a_i/v)\big] \gtreqless 0 \quad \text{as } p_m a_i \gtreqless v$$

so optimal profit under noncompliance is

$$\pi_i^n = \begin{cases} A_i\big[p_m(a_i + b_i) \\ \quad - v - v\ln(p_m a_i/v) - c_i\big], & \text{if } p_m a_i > v \\ A_i(p_m b_i - c_i) & \text{otherwise.} \end{cases}$$

Now consider behavior under compliance with a voluntary government program such as the wheat and feed grain programs that employ the following policy instruments: a target price p_t, a support price p_s, a diversion requirement λ, a government payment per acre for diversion g, and a program yield y_p. Define $p^* = \max(p_s, p_m)$. Then, under the provisions of these programs whereby farmers are entitled to at least the target price on production up to the program yield, and to at least the support price on all their production, profit for farm i is

$$\pi_i = A_i(1 - \lambda)\big[\max(p_t, p_m)\cdot y_p + p^*\cdot\max(y_i - y_p, 0) - vx_i - c_i\big] + A_i\lambda g.$$

The first-order condition for profit maximization is $\partial\pi_i/\partial x_i = A_i(1 - \lambda)(p^* a_i e^{-x_i} - v) = 0$, if $y_i > y_p$ and $\partial\pi_i/\partial x_i = A_i(1 - \lambda)(-v) < 0$ if $y_i < y_p$. Thus, the optimal input use per acre under compliance is $x_i^c = \ln(p^* a_i/v)$ if $p^* a_i > v$ and $y_p < a_i + b_i - v/p^*$, and 0 otherwise.

The optimal profit under compliance is

$$\pi_i^c = \begin{cases} A_i(1-\lambda)[p_t y_p - c_i] + A_i\lambda g & \text{if } y_p \geq a_i + b_i - v/p^* \\ A_i(1-\lambda)\big[(p_t - p^*)y_p + p^* b_i - c_i\big] \\ \quad + A_i\lambda g & \text{if } y_p < b_i,\ p^* a_i \leq v \\ A_i(1-\lambda)\big[(p_t - p^*)y_p + p^*(a_i + b_i) \\ \quad - v - v\ln(p^* a_i/v) - c_i\big] + A_i\lambda g & \text{otherwise.} \end{cases}$$

Comparing π_i^c and π_i^n, it is clear that compliance can be induced by a sufficiently high target price and diversion payment regardless of how low the support price is and how high the diversion requirement is. On the other hand, more intense use of the polluting input is not induced by a high target price and diversion payment if the program yield is sufficiently high, and/or the support price is sufficiently low. For example, more intense use is not induced if the support price is at or below the market price, and the program yield is not below the yield that occurs with market price under noncompliance.

The important point of these results is that current agricultural policies are sufficiently general to permit independent control of the extensive and intensive margins. The extensive margin, (i.e., the choice of diverted and producing acreage) can be controlled through the target price, the diversion requirement, and the diversion payment. The intensive margin can be controlled within certain bounds by the choice of support price and program yield.

V. Conclusions

There are a number of directions in which the framework presented here needs to be extended. The disaggregate model needs to be aggregated so that market equilibrium implications can be analyzed. Dynamic aspects of these issues need to be considered, especially the implications for investment behavior. Empirical research must address the technical detail involved in integrating physical and economic models.

Both agricultural production and environmental impacts depend on highly location-specific environmental conditions. Reality is far too complex to allow generalizations about the environmental impacts of agricultural policies. Our analysis points to the kinds of data that are needed to make valid inferences. Statistically reliable field-specific production data and environmental data would make possible measurement of key parameters (such as the correlation between production decisions and environmental attributes of land) that are needed to assess the aggregate relationships between agricultural policy, environmental policy, and the environment.

REFERENCES

Antle, J. M. and Just, R. E., "Effects of Commodity Program Structure on Resource Use and the Environment," in N. Bockstael and R. Just, eds., *Commodity and Resource Policy in Agricultural Systems*, New York: Springer-Verlag, forthcoming 1990.

Hochman, E. and Zilberman, D., "Examination of Environmental Policies Using Production and Pollution Microparameter Distributions," *Econometrica*, July 1978, *46*, 739–60.

National Research Council, *Alternative Agriculture*, Washington: National Academy Press, 1989.

Coordinating Agricultural and Environmental Policies: Opportunities and Tradeoffs

By S. R. JOHNSON, ROBERT WOLCOTT, AND SATHEESH V. ARADHYULA*

The general policy framework for agriculture in the United States dates to the New Deal era of the 1930s (Gordon Rausser, 1990). Since then, the U.S. government has maintained policies to stabilize domestic markets for major program crops and maintain farm income. These policies have also addressed the orderly marketing of specialty crops through an array of supply management, processing, and distribution measures. With the advent of the Soil Conservation Service in 1935, the U.S. Department of Agriculture (USDA) became directly involved in "voluntary" environmental policy. In fact, voluntary participation in programs to manage commodity markets and aggregate farm income, and the environment, has been a hallmark of the general policy approach of the USDA.

Management of the agricultural sector and the environment with multiple-objective national policies and voluntary participation implies a number of administrative and targeting problems. These range from conflicts in the objectives of environmental and agricultural policies to more technical issues of incidence, free ridership, and efficiency. The advent of the Food Security Act of 1985 signaled a significant change in USDA agricultural and environmental policy. Eligibility for participation in commodity and other federal agricultural programs was made conditional on future compliance with farming practices limiting soil loss (conservation compliance). A conservation reserve was created for highly erodible land. The latter was in some respects similar to the soil bank of the 1960s, but targeted more specifically to soil conservation. Other provisions for limiting the agricultural use of frail lands were also included (Sandra Batie and J. P. Marshall, 1989).

This change in USDA policy has broadened opportunities for coordinating environmental and agricultural price stabilization and income maintenance measures, and interests in better identifying associated opportunities and tradeoffs. In the 1990 Farm Bill debate, environmental and food safety titles are receiving increased attention (Charles Benbrook, 1989). Water quality has become a national concern (EPA, 1987). There is more direct intervention by the EPA and other federal agencies responsible for natural resources (EPA). Moreover, states have become important players in the management of the environment. The result is an increasingly complex set of agricultural and environmental policies with opportunities for complementarities and tradeoffs that are difficult to identify and estimate, but with potential for coordination to improve their efficiency and focus.

I. Opportunities and Tradeoffs

Opportunities and tradeoffs for agricultural and environmental policies are conditioned by the economic structure of the agricultural sector, resource and technological endowments, and resource markets. At first, this would seem a simple enough observation. The complicating factor is that the economic structure of the sector and the resource markets are in turn conditioned by current and past policies. These policies extend beyond apparent regulatory measures to technology development and institutional design. The policies are themselves complex, much more so than we would conclude by reviewing the results of stylized economic analyses. Policies as usually conceived (for

---

*Johnson and Aradhyula: Center for Agricultural and Rural Development, Department of Economics, Iowa State University, Ames, IA 50011; Wolcott: Environmental Protection Agency, Washington, D.C.

example, agricultural commodity programs, water quality standards, etc.) are in fact packages of instruments with specific structures that influence possibilities for achieving objectives for environmental and agricultural constituents.

To illustrate these observations, first consider how current and past policies for agriculture and the environment influence opportunities and tradeoffs for improved coordination. Examples of associated relationships are abundant. Base acreages and associated subsidies for commodity programs are capitalized into land values; crop insurance that is not actuarily fair influences cropping patterns by affecting producer risks; state technical guidelines determine land eligible for the conservation reserve; and registration decisions for pesticides change crop production costs. In short, policies have distributional and rent-creating features that alter relative prices and the structure of agricultural and resource markets. In the longer term, these policies also influence technical change and even available resource stocks.

Second, there is the broader question of institutional design and the political economy of agricultural and resource policy. Here opportunities and tradeoffs appear less well understood and researched (Johnson, 1985; Rausser). Patent laws for agricultural chemicals and seeds which shift research expenditures from public institutions to the private sector, property rights of landowners for mining water and soil resources, and technology and policy outcomes that create markets and more efficient exchange (for example, options markets for agricultural commodities), all influence possibilities for policy action and the tuning of existing policies. Changing political organization (such as the representation of agricultural constituencies and commodity check-off referenda) and the availability of new information (in chemical residuals in groundwater and the food chain) also condition the institutional setting for agricultural and environmental policy, and associated opportunities and tradeoffs.

Third, there is the structure of policies, and the fact that programs for agriculture and the environment are "packages" of instruments organized according to a specific design. The programs available, the instruments they incorporate, and their structures influence possibilities for coordination to improve efficiency in meeting intended policy objectives. Possibilities for change within the context of existing programs can be viewed as "tuning." Economists have to date tended to stylize program specifications in policy analyses, and it would appear underestimate the importance of program design in evaluating agricultural and environmental regulations. The bureaucratic and political interests that determine policy, along with economic outcomes, play out within the structure of the programs and thus afford specialized opportunities and tradeoffs.

An analogy to "macros" in computing is tempting in assessing opportunities and tradeoffs for environmental and agricultural policy. At one level, we can manipulate the programs (macros) to achieve desired outcomes. But, at this level, the opportunities and tradeoffs are constrained by the structures of the programs and the instruments they make available. New opportunities and tradeoffs are provided by the new macros, and tuning and changes of structure for existing programs.

The present analysis of tradeoffs and opportunities through coordination of agricultural policy has quite modest objectives. The intent is to illustrate tradeoffs possible within a particular agricultural program. Clearly, the results are influenced by other policies, the institutional design, and the current structure of agricultural commodity programs. The illustration, however, provides an initial set of results on opportunities and tradeoffs, and perhaps a glimpse of the richness that awaits more detailed consideration of policy structure as an element of policy change.

II. An Illustrative Example

The model used for the illustrative analysis integrates output supply and input demand in a consistent behavioral framework (Aradhyula and Johnson, 1990). A normalized quadratic profit function was estimated, with the instruments, architecture, and functioning of commodity programs structurally described. The model estimates output supply of thirteen major agricultural commodities with four variable input demands (land, fertilizer, operating capital, and labor) simultaneously using full-information maximum likelihood methods. Annual USDA aggregate time-series data for the United States were used, 1950–86. Homogeneity, symmetry, and convexity of the restricted profit function were imposed.

The parameters of the estimated model yielded plausible own-price elasticities, but cross-price elasticities smaller than from previous studies with less detailed specifications for commodity programs and fewer commodities. Participation rates in programs for wheat, rice, corn, other coarse grains, and cotton were estimated using a structural policy specification external to the profit function (Aradhyula-Johnson).

The relevant equations from the profit function formulation for the ith output supply are

$$Y_i = a_i + \sum_{j=1}^{16} b_{ij}P_j + \sum_{k=1}^{2} c_{ik}Z_k \tag{1}$$

$$P_i = PAR_i PP_i + (1 - PAR_i)NP_i \tag{2}$$

$$PP_i = \max(TP_i, FP_{it-1}) \tag{3}$$

$$NP_i = FP_{it-1} \tag{4}$$

$$P_{14} = PAR_i RENT/(1 - SET_i) + (1 - PAR_i)RENT \tag{5}$$

$$PAR_i = 1 - \exp[\alpha_{i1}(TP_i/FP_{it-1}) \times (1 - SET_i - DIV_i) + \alpha_{i2}(PDP_i/FP_{it-1})DIV_i] \tag{6}$$

where Y_i is the output supply level for the ith community, P_i is aggregate price, TP_i is target price, FP_i is farm price, PAR_i is commodity program participation rate, PP_i is participant price, NP_i is nonparticipant price, SET_i is set-aside requirement, DIV_i is paid-diversion level, PDP_i is paid-diversion payment rate, P_{14} is aggregate rental value of the land, $RENT$ is rental rate for land, Z_1 is quantity of fixed capital, Z_2 is a time variable, and a_i, b_{ij}, c_{ik}, α_{i1}, and α_{i2} are estimated parameters. Variables without a time subscript refer to current period.

Program Structure. To analyze the effects of a change in target price of ith output on the own-production level (1) can be differentiated with respect to TP_i to obtain

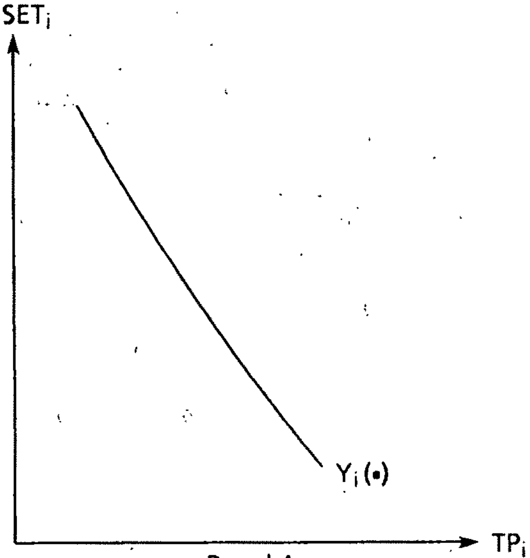
$$\begin{aligned} \partial Y_i/\partial TP_i &= b_{ii}[(PP_i - NP_i)(\partial PAR_i/\partial TP_i) + PAR_i(\partial PP_i/\partial TP_i)] \\ &\quad + b_{i14}[RENT\{SET_i/(1 - SET_i)\} \times (\partial PAR_i/\partial TP_i)]. \end{aligned} \tag{7}$$

Given that α_{i1} and α_{i2} are negative, the first term on the right-hand side of (7) is nonnegative and reflects the output expansion effect of the target price. For normal technology, $b_{i14} < 0$, and the second term on the right-hand side of (7) is negative. The negative effect of an increase in target price on own-output supply is due to increased program participation, and the associated acres set aside. Thus, from equation (7), a change in the target price will have two effects on own output, acting in opposite directions.

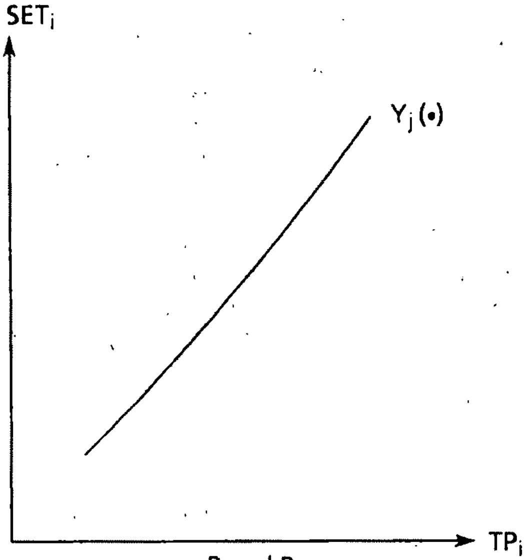
The effect of a change in the set-aside requirement (SET_i) on own-output supply is similarly obtained:

$$\begin{aligned} \partial Y_i/\partial SET_i &= b_{ii}[(PP_i - NP_i)(\partial PAR_i/\partial SET_i)] \\ &\quad + b_{i14}\Big[PAR_i RENT/(1 - SET_i)^2 + RENT\{SET_i/(1 - SET_i)\} \times (\partial PAR_i/\partial SET_i)\Big] \end{aligned} \tag{8}$$

It can also be shown that a change in the set-aside requirement also has separate and opposite effects on own-output supply. Thus, added information is required to determine the exact output effect.



Panel A
Iso-output curve for corn.



Panel B
Iso-output curve for soybeans.

FIGURE 1. TRADEOFFS BETWEEN THE TARGET PRICE ON SET-ASIDE REQUIREMENT FOR CORN

Selected tradeoffs between policy instruments implied by the estimated model are presented in Figure 1. The two policy instruments chosen for illustration are the target price and the set-aside requirement for corn. Panel A in Figure 1 illustrates the iso-output curve in *SET* and *TP* space with slope,

$$(9)\quad dSET_i/dTP_i = -(\partial Y_i/\partial TP_i)/(\partial Y_i/\partial SET_i).$$

For this iso-output curve, expression (9) was evaluated at 1986 sample values for all variables, using the estimated parameters. The iso-output curve in Panel B was obtained similarly but for soybean, a commodity with a different government program but highly substitutable for corn in current production systems. The slope of the iso-product curve for soybean is

$$(10)\quad dSET_i/dTP_i = -(\partial Y_j/\partial TP_i)/(\partial Y_j/\partial SET_i),$$

where the subscripts i and j denote corn and soybean, respectively. Using equations (1)–(6), it can be shown that the expression in (10) is positive. Higher target prices and lower set-asides for corn move acreage toward corn resulting in a lower soybean output. This cross effect of the corn program on soybean output indicates the complex nature of the policy structure. But, such interdependencies in output supply present opportunities for coordinating agricultural and environmental policies.

External and Program Impacts. Table 1 presents estimated impacts of changes in selected policy instruments and external factors. Estimates for only two inputs (fertilizer and labor) and two outputs (corn and soybeans) are presented due to space limitation. Policy instruments evaluated are target prices and set-aside requirements for corn. External factors are input price taxes. All impacts are presented as elasticities and for 1986. Results indicate that changes in the program instruments, target prices and set-aside requirements, had positive effects on own-output supply. The effect of changes in *TP* and *SET* for corn on soybean output was small in magnitude. Increased *TP* of corn had a positive effect on the aggregate input demand for labor while it appeared to exert a negative effect on aggregate fertilizer use.

TABLE 1—ESTIMATED IMPACTS OF CHANGES IN SELECTED PROGRAM INSTRUMENTS AND EXTERNAL FACTORS

	Input Demand		Output Supply	
Policy Change	Fertilizer	Labor	Corn	Soybean
Commodity Program				
Target Price for Corn	−0.14	0.05	0.04	−0.02
Set-Asides for Corn	0.04	0.04	0.16	0.01
Input Price Tax				
Land Rent	−0.01	0.31	−0.01	0.02
Fertilizer Price	−0.18	0.21	0.01	−0.02
Operating Capital Price	0.05	0.86	−0.01	−0.08
Wage Rate	0.88	−1.73	−0.06	−0.05

Notes: All values are estimated for the year 1986.

The aggregate nature of the input demand specifications to an extent explains the small magnitudes of the elasticities. Similar estimates for the set-aside requirement for corn on aggregate input demand were obtained. A 1 percent increase in set-aside requirement for corn increased aggregate fertilizer use by 0.04 percent.

Aggregate input demand functions were more responsive to factor prices. Of all inputs, labor demand was the most sensitive. Moreover, all input demands were more sensitive to wage rate than other factor prices. It is emphasized that the elasticities were evaluated for 1986, and would change depending on conditioning values, including technology. Generally, the results in Table 1 show that even crop-specific program instruments have sector-wide influences on output and input use. The input price changes reflecting taxes were included to suggest possible environment tradeoffs. Of course, the proper way to estimate such impacts is with equally structured representations of environmental programs.

III. Concluding Observations

The complexity of identifying opportunities and tradeoffs for agricultural and environmental policy has been sketched. Program structure and the institutional setting were emphasized as important conditioning factors. Results from the example illustrated how tradeoffs implied by program structure can be evaluated. Analysis can be made richer by more fully incorporating program structure, and extend information on possibilities for coordinating agricultural and environmental policy.

REFERENCES

Aradhyula, Satheesh V. and Johnson, Stanley R., "Policy Structure, Output Supply and Input Demand for U.S. Crops," CARD Working Paper, Iowa State University, 1990.

Batie, Sandra S. and Marshall, J. P., *A Proposed New Structure for Food and Agricultural Policy*, Blacksburg: VPI and SU Information Series 1989-1, July 1989.

Benbrook, Charles M., "Agriculture and Groundwater Quality: Policy Implications and Choices," American Association for the Advancement of Science, San Francisco, 1989.

Johnson, Stanley R., "A Critique of Existing Models for Policy Analysis," *Proceedings, Agricultural Sector Models for Policy Analysis*, Agriculture Canada, Ottawa, 1985.

Rausser, Gordon C., "The Evolution and Coordination of U.S. Commodity and Resource Policies," in N. Bockstael and R. E. Just, eds., *Commodity and Resource Policies in Agricultural Systems*, New York: Springer-Verlag, forthcoming 1990.

U.S. Environmental Protection Agency (EPA), "Agricultural Chemicals in Ground Water: Proposed Pesticide Strategy," Office of Pesticides and Toxic Substances, Washington: USGPO, 1987.

Limitations in Evaluating Environmental and Agricultural Policy Coordination Benefits

By James Hrubovcak, Michael LeBlanc, and John Miranowski*

Agricultural policy has traditionally been concerned with the maintenance and stability of farm income, and the provision of a stable food supply at low relative prices. The Food Security Act of 1985 was the first farm legislation to directly tie farm income and environmental concerns through the Conservation Reserve Program, conservation compliance, sodbuster, and swampbuster provisions. The coordination of conservation and farm income policies contributed to commodity supply control by reducing the availability of arable land, and it promoted environmental quality by removing some highly erodible land and wetlands from crop production.

A number of new environmental and food safety initiatives are also being enacted outside the agricultural policy process. The most significant piece of legislation is the Water Quality Act of 1987, designed to control agricultural nonpoint source pollutants that affect water quality. In addition, amendments to the Federal Insecticide, Fungicide, and Rodenticide Act are being considered. Efforts are also underway to protect groundwater and endangered species.

Issues like food safety and water quality have brought agriculture and nonpoint source pollution to the forefront of environmental attention. Growing environmental awareness emphasizes the emerging importance of nontraditional agricultural policy concerns. Future environmental and food safety policies, are likely to affect agriculture and greater coordination between farm and environmental policy is likely to assure more satisfactory outcomes.

Because agricultural chemicals are central to many environmental and food safety concerns, we illustrate some of the difficulties of coordinating agricultural and environmental policies by altering agricultural chemical use. Reductions in the use of agricultural chemicals may be achieved by price incentives (i.e., taxes on use, subsidies for reduced use), regulations and direct controls (i.e., registration, bans), improved information on control alternatives, and risk reduction strategies (i.e., insurance). For ease in illustrating the limitation of policy coordination, we consider only taxes and direct controls. Although other considerations are relevant, our example highlights the central importance of consistent behavioral incentives provided by policies and adequate knowledge of the underlying agricultural production structure for coordination of policy.

I. A Framework for Evaluating Policy

Controlling chemical use in agricultural production provides an excellent example of the limitations in evaluating agricultural, environmental, and food safety policy coordination. Significant crop yield increases over the last several decades have been associated with the adoption of pesticides and fertilizers. Furthermore, government supply control measures may have led to more intensive use of agricultural chemicals (Miranowski et al., 1990). At the same time, agricultural chemicals may impose economic costs on the environment and human health. Such tradeoffs need to be weighed when assessing the benefits and costs from integrating agricultural, environmental, and food safety policies.

To assess these tradeoffs, we employ a computable general equilibrium model. Our use of the term general equilibrium corresponds to the well-known Arrow-Debreu model (discussed in detail in Kenneth Arrow and F. H. Hahn, 1971). The general equilibrium nature of the model is reflected by the determination of price vectors for consumer and producer goods and services that clear all markets. Equilibrium prices determine the optimal allocation of resources, given the endowments of labor, capital, and natural resources.

---

*Economic Research Service, U.S. Department of Agriculture, Washington, D.C. 20005. The views expressed herein do not necessarily reflect those of the U.S. Department of Agriculture.

The model presented here is explained in Roy Boyd (1987). The model is a static computable general equilibrium model that includes 12 production sectors. Agriculture is decomposed into program crops, livestock and dairy, and all other agricultural production. A separate constant elasticity of substitution production function specifies the technology for each sector. Land is an input in forestry and agriculture.

Six representative consumers, differentiated by income, supply all primary factors. Each household's consumption is modeled as a triple-nested utility function. Commodity market demands depend on all prices, are continuous, nonnegative, homogeneous of degree zero, and satisfy Walras' Law. A government sector is included to collect taxes and return revenue back into the economy. A foreign consumer is included to close the model. Full employment of all primary factors of production is assumed and the government budget and foreign trade accounts are balanced.

II. Agricultural Impacts of Taxing Agricultural Chemical Use

A tax on agricultural chemicals increases the "effective" price of chemicals and reduces use. Higher chemical prices increase production costs, increase commodity prices, and reduce overall output (Table 1). Crop production decreases more than livestock production, because chemicals represent a larger share of production expenses. Lower agricultural output reduces the demand for all inputs including land. The increase in the relative price of chemicals, however, also stimulates the demand for land. Assuming a moderate degree of input substitution (Allen elasticity between land and chemicals is 0.7), land rents increase as the tax on chemicals increases.

In a static and certain world, the environmental damages caused by agricultural chemical use could be reduced by setting the appropriate tax on chemicals. However, increases in productivity combined with shifts in foreign demand create uncertainty with respect to future economic conditions, and lead to problems coordinating commodity and environmental policies.

TABLE 1—IMPACTS OF TAXING AGRICULTURAL CHEMICAL USE

	Percent Tax Rate				
	10	25	50	75	100
Quantities					
Program crops	−1	−2	−3	−4	−5
Livestock	0	0	−1	−1	−2
Nonprogram crops	−1	−1	−2	−3	−4
Chemical use	−6	−13	−22	−29	−34
Prices					
Program crops	1	2	5	7	8
Livestock	1	1	2	3	4
Nonprogram crops	1	3	5	7	9
Land rents	1	2	4	6	7

Note: Percentage change from base.

While the benefits of commodity and environmental programs often overlap, the goals of each program differ. A basic goal of commodity programs is to support income and stabilize prices. Currently, this is accomplished by price supports on agricultural production coupled with acreage set-aside requirements to control supply by removing land from production. Set-aside requirements coupled with higher effective prices for agricultural output could lead to more intensive input use on the land remaining in production. Greater chemical use, as a substitute for land, may lead to greater environmental and food safety concerns.

Changes in price supports or output prices can alter the effectiveness of a chemical tax. This can be illustrated by altering the level of price supports for program crops, modeled here as an ad valorem subsidy. The tax on agricultural chemicals is fixed at 50 percent. An increase in price supports or an increase in commodity prices due to an expansion in export demand can reduce the effectiveness of a chemical tax by increasing the opportunity cost of reducing chemical use (Table 2).

Alternatively, reducing the level of price supports with a constant chemical tax leads to declines in both agricultural output and

TABLE 2—EFFECTS OF A TAX AND DIRECT CONTROLS UNDER ALTERNATIVE LEVELS OF FARM PRICE SUPPORTS

	Percentage Change in Support				
	+15	+5	0	-5	-15
Taxes					
Quantities					
Program crops	3	-1	-3	-5	-10
Livestock	-1	-1	-1	-1	-1
Nonprogram crops	-4	-3	-2	-2	-1
Chemical use	-17	-20	-22	-23	-27
Market Prices					
Program crops	-6	1	5	9	20
Livestock	5	3	2	2	0
Nonprogram crops	8	6	5	4	2
Land rents	21	10	4	-1	-11
Direct Controls					
Quantities					
Program crops	2	-1	-3	-5	-9
Livestock	-2	-1	-1	-1	0
Nonprogram crops	-4	-3	-2	-2	-1
Chemical use	-22	-22	-22	-22	-22
Market Prices					
Program crops	-4	1	5	8	17
Livestock	6	3	2	1	0
Nonprogram crops	9	6	5	4	1
Land rents	22	10	4	-1	-12

Note: Percentage change from base. Changes in support are introduced through a unit output subsidy or tax. (+) indicates a greater farm support, (−) indicates less farm support.

TABLE 3—EFFECTS OF A TAX AND DIRECT CONTROLS UNDER ALTERNATIVE SUBSTITUTION ELASTICITIES

	Scenario				
	A	*B*	*C*	*D*	*E*
Taxes					
Quantities					
Program crops	-2	-3	-3	-3	-3
Livestock	0	-1	-1	-1	-1
Nonprogram crops	-2	-2	-2	-2	-3
Chemical use	-7	-13	-22	-29	-38
Prices					
Program crops	3	4	5	5	5
Livestock	1	2	2	3	3
Nonprogram crops	4	4	5	5	5
Land rents	-3	1	4	5	7
Direct Controls					
Quantities					
Program crops	-8	-5	-3	-2	-2
Livestock	-2	-1	-1	-1	-1
Nonprogram crops	-7	-4	-2	-2	-1
Chemical use	-22	-22	-22	-22	-22
Prices					
Program crops	13	8	5	3	2
Livestock	4	3	2	2	1
Nonprogram crops	17	9	5	4	3
Land rents	-13	1	4	4	4

Note: Percentage of change from base. Allen partial elasticities of substitution are σ_{ac} = land/chemicals, σ_{lc} = labor/chemicals, and σ_{kc} = capital/chemicals.

A: $\sigma_{ac} = 0.175$, $\sigma_{lc} = \sigma_{kc} = 0.125$.
B: $\sigma_{ac} = 0.35$, $\sigma_{lc} = \sigma_{kc} = 0.25$.
C: $\sigma_{ac} = 0.70$, $\sigma_{lc} = \sigma_{kc} = 0.50$.
D: $\sigma_{ac} = 1.05$, $\sigma_{lc} = \sigma_{kc} = 0.75$.
E: $\sigma_{ac} = 1.50$, $\sigma_{lc} = \sigma_{kc} = 1.00$.

chemical use. The complementary relationship between chemical taxation and reductions in farm price supports suggests that reducing price supports would allow a reduction in the chemical tax rate while maintaining the desired reduction in chemical use. However, the uncertainty with respect to future levels of price supports makes it difficult to set the appropriate tax rate, *ex ante*.

Direct control of agricultural chemicals is an alternative to taxes which insulates policies from changes in farm support or output prices. Controls can be used to set the maximum amount of chemicals applied. Direct controls can be set to reproduce the effects of a 50 percent tax (Table 2), and can be interpreted as a state contingent tax on chemicals where the tax rate on chemical use varies with the level of farm support.

One problem with state contingent taxes is high transition costs. The preponderance of fixed and quasi-fixed factors in agriculture suggests it is unlikely that the short-run production technology in agriculture is sufficiently flexible to respond to constantly changing tax rates. Changes in tax rates, therefore, would be reflected in farm income changes rather than chemical use.

Inadequate or incorrect information regarding the structure of agricultural production can also lead to incorrect or inaccurate policy prescriptions. This uncertainty is represented by varying the substitution possibilities between chemical and nonchemical inputs (Table 3). Farm programs are assumed intact, and the substitution elasticities between chemical and nonchemical inputs are varied. Different assumptions regarding the substitution possibilities lead to significant differences in the effect of taxes and direct controls on chemical use, agricultural output, prices, and who ultimately bears the burden of the tax.

If agricultural production were relatively chemical intensive, determining the incidence of the tax is straightforward; chemicals suppliers will share a relatively larger burden of the tax through reduced demand and lower prices. However, because agricultural production is land intensive, the burden of the tax depends on the substitutability between inputs. A less flexible production technology coupled with our assumption of full employment of land generates a smaller decrease in chemical use and in agricultural output. With limited substitution, the burden of the tax is shifted, in part, from consumers and chemical suppliers to land owners.

Under a more flexible production technology, the opposite is true. Chemical use is more responsive to a change in the tax rate, resulting in larger output and price impacts. A 50 percent tax leads to a large decrease in chemical use. Because nonchemical inputs are more easily substituted for chemicals, the demand for farmland increases, causing land rents to increase and the tax burden to be shifted to consumers and chemical suppliers.

With direct controls, chemical use is unaffected by the degree of substitution between inputs, but output and price effects are different than under a tax. Unlike the tax, because chemical use must decline with direct controls, output decreases and price increases are larger when substitution possibilities are limited. Similar to the tax, land rents also decline with limited substitution. In this case, incorrect assumptions concerning substitution possibilities have more serious consequences for consumers, chemical suppliers, and land owners than would a tax.

III. Conclusions

Although our illustration highlights two key problem areas associated with coordinating environmental and agricultural policy, at least three additional points merit discussion. Policy effectiveness depends on whether policy changes are temporary or permanent or whether the policymaker operates with rules or uses discretion (Robert Barro and David Gordon, 1983). Important elements include the tension between static and dynamic efficiency, time inconsistency (Finn Kydland and Edward Prescott, 1977), and changes in the underlying economic structure (Robert Lucas, 1976). Second, while the efficiency and distribution dimensions of policy effectiveness have been widely discussed in the literature, the certainty issue needs greater attention. Because there are many targets and instruments in agricultural resource policies, and because the response of target variables to policy actions is uncertain, an optimal policy would employ all instruments (William Brainard, 1967). Finally, many resource and agricultural problems are long term and require a dynamic framework to assure analytical consistency.

Given the assumptions and model structure used in our analysis, the results and discussion indicate that public policies designed to simultaneously satisfy farm income and environmental objectives face some serious challenges. Efforts to achieve a reduction in agricultural chemical use through taxes should impact chemical use and reduce environmental residuals. But output price and production uncertainties, coupled with uncertainties about the elasticities of substitution between key inputs, generate significant uncertainties on the beneficial environmental impacts.

Direct controls on agricultural chemical use may reduce the perceived uncertainty of farmer response, but implementing and enforcing such controls introduces a host of problems. Also, inadequate knowledge regarding the underlying production structure may lead to assumed elasticities of input substitution that overstate or understate the output and price effects of direct controls, even though the environmental impacts may be more certain.

REFERENCES

Arrow, Kenneth J. and Hahn, F. H., *General Competitive Analysis*, San Francisco, 1971.

Barro, Robert J. and Gordon, David B., "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, July 1983, *12*, 101–12.

Boyd, Roy, "An Economic Model of Direct and Indirect Effects of Tax Reform on Agriculture," Technical Bulletin No. 1743, USDA Economic Research Service, 1987.

Brainard, William, "Uncertainty and the Effectiveness of Policy," *American Economic Review Proceedings*, May 1967, *57*, 411–25.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conference Series on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

Miranowski, John A., Hrubovcak, James M. and Sutton, John D., "The Effects of Commodity Programs on Resource Use," in Nancy Bockstael and Richard Just, eds., *Commodity and Resource Policy in Agricultural Systems*, New York: Springer-Verlag, forthcoming.

BARGAINING AND PRICE FORMATION UNDER INCOMPLETE INFORMATION: THEORIES AND EXPERIMENTS[†]

Explicit Communication and Bargaining Outcomes

By VINCENT P. CRAWFORD*

Bargaining is a pervasive phenomenon in modern economies, and the leading examples of labor negotiations, trade agreements, and strategic arms limitation talks suggest that the potential welfare gains from improving the efficiency of bargaining outcomes are enormous. In recent years, there has been significant progress toward the kind of theoretical and empirical understanding of bargaining needed to realize some of those gains. However, in most of this work, it is assumed that bargaining is *tacit* in the sense that bargainers can communicate only by making offers and counteroffers that directly affect their payoffs. Real bargaining, by contrast, is usually *explicit*, in that bargainers can also communicate by sending nonbinding messages with no direct payoff implications. Such "cheap talk" messages evidently play an important role in coordinating bargainers' expectations so that they can reach agreement, and in determining how they share the resulting surplus.

This paper discusses theoretical and experimental work on explicit bargaining. Section I reviews the theory of tacit bargaining; Section II discusses the theory of explicit bargaining; and Section III reviews the experimental evidence on communication and bargaining outcomes, and discusses the gaps between theory and evidence that remain.

I. Tacit Bargaining

The modern theory of bargaining originated in John Nash's (1953) analysis of the bargaining problem faced by two fully informed players who have an opportunity to meet before playing a game and make a binding agreement about how to play it. Nash's innovation was to construct a complete model of the players' interaction, including a model of the bargaining process, and then to use his notion of equilibrium to characterize their rational responses to its possibilities. He followed the pattern of his 1950 axiomatic bargaining analysis in summarizing the bargaining environment by the set of utility combinations associated with feasible agreements about strategy choices in the underlying game, here called the *utility-possibility set*, and players' utilities when they do not reach agreement, here called their *disagreement utilities*. In that analysis, he sought to express the utilities that bargainers could rationally "anticipate" as a function of the utility-possibility set and their disagreement utilities, and showed that one such function (now called the Nash bargaining solution) is uniquely characterized by four plausible axioms. These axioms extend the widely accepted principle of equally sharing the gains from agreement to bargaining problems with nonlinear utility-possibility frontiers and arbitrary disagreement outcomes, with the gains measured relative to disagreement.

In Nash's (1953) bargaining model, bargainers make simultaneous *demands*, expressed in utility terms. If their demands are *compatible* in the sense that they lie in the utility-possibility set, then bargaining ends with a binding agreement about how they will play the underlying game that yields them the utilities they demanded. Otherwise, bargaining ends with no agreement and bargainers get their disagreement utilities, here assumed exogenous. Nash observed that any pair of demands associated with an efficient agreement that is at least as good as disagreement for each bargainer is an equilibrium in this game: If either bargainer made a lower demand, his payoff would be reduced with no compensating benefit; and if he made a higher demand, disagreement would result. Thus, although there are normally many such agreements, all are consistent with equilibrium (and its standard refinements) and, even when there is no uncertainty about the environment, the bargaining game is likely to generate a great deal of uncertainty about how players will respond to its multiplicity of equilibria. Unless bargainers find a way to resolve this *strategic uncertainty*, they may not realize any of the gains from reaching an agreement: At the heart of the bargaining problem lies a coordination problem.

[†] *Discussants*: Robert Anderson, University of California-Berkeley; Preston McAfee, University of Western Ontario and California Institute of Technology; William Samuelson, Boston University.

*Department of Economics, University of California-San Diego, La Jolla, CA 92093. I am grateful to the NSF for research support.

Nash (1953) suggested two complementary resolutions of this coordination problem. In one, he noted that the normative force of his 1950 bargaining solution might focus bargainers' expectations on the associated demand-game equilibrium. In the other, he outlined a "smoothing" argument, showing that his solution is the "only necessary limit" of the equilibria in smoothed demand games as the amount of smoothing approaches zero.

A third resolution is due primarily to Ariel Rubinstein (1982). He replaced Nash's simultaneous-offers model with a sequential-offers model, in which fully informed bargainers take turns making offers to each other (with one exogenously assigned to the right to make the first offer). A bargainer can accept his partner's latest offer at any stage, in which case it becomes a binding agreement, taking effect immediately. But, a bargainer who rejects an offer must wait until the next stage before making a counteroffer. This process continues indefinitely until an offer is accepted; but delay is costly, in that bargainers prefer to reach agreement sooner rather than later.

Rubinstein showed that this model has a unique subgame-perfect equilibrium even when the time horizon is infinite: When bargainers are fully informed, his assumptions about the rules of bargaining completely eliminate the coordination problem. In this equilibrium, the bargainer who makes the first offer chooses it so as to extract all of his partner's surplus from accepting it, taking into account that the alternative is to make a counteroffer, chosen in the same way, but one stage later. In equilibrium, this first offer is always accepted, so the outcome is always efficient; and how bargainers share the resulting surplus in it is completely determined by the relationship between their delay costs. Although the right to make the first offer is an advantage, in the leading case in which bargainers discount the future at the same constant rate, this advantage disappears as their discount factors approach one (or as the time between offers approaches zero) and the unique subgame-perfect equilibrium outcome then approaches Nash's bargaining solution.

The choice between simultaneous-offers and sequential-offers specifications has important implications for the role of explicit communication. Although models like the demand game strike many people as too simple to reflect the dynamics of real bargaining, there is more to this choice than whether bargaining is static or dynamic. Suppose, following Thomas Schelling (1960, Appendix B) and John Harsanyi and Reinhard Selten (1988, pp. 23–26), that there is a fixed, known deadline, and that the only rules of bargaining are the stipulations that, if an agreement is reached by the deadline, it will be enforced, and, if not, the process ends in disagreement. (This rules out artificial constraints on the timing of offers like those in Rubinstein's model, but allows symmetric, physical constraints on the time it takes to make or transmit offers.) Then, if the costs of delay are negligible, for reasonable specifications of bargainers' strategic possibilities, there is an equilibrium that correspond to each of the efficient equilibria in Nash's demand game.

More generally, even with no fixed deadline, bargainers' strategies are dynamic decision rules that can be viewed as chosen simultaneously, and in any equilibrium in which bargainers are certain to reach agreement, each bargainer's strategy must be such that the lowest utility his partner can make him accept is just compatible with the lowest utility he can make his partner accept. These lowest utilities function just as the demands in Nash's model, and the resulting agreements correspond precisely to its efficient equilibria. Thus, bargaining can pose a significant coordination problem even when bargainers are fully informed and approach agreement by a sequence of alternating offers. The give-and-take of real bargaining is surely, at least in part, a robust response to this coordination problem.

The choice between specifications therefore turns on the plausibility of the rules in sequential-offers models, and on whether the costs of delay are more important than the risk of coordination failure. In most contemplated applications, bargainers are free to choose the timing of their offers and counteroffers, within wide limits, and the costs of delay are small compared to the risk of coordination failure. It seems clear that this risk is not adequately modeled by the common practice of viewing delay costs as reflecting the chance that bargaining will be exogenously terminated before an agreement is reached.

There is also substantial experimental evidence that coordination is the essence of the bargaining problem, and that focal points based on the structure of the game, the context in which it is played, or the history of play in analogous games, as discussed by Schelling, strongly influence bargaining outcomes. Alvin Roth (1985) and his collaborators found, in environments designed to conform to the assumptions of models like those of Schelling and Harsanyi-Selten just discussed, that subjects' agreements were heavily influenced by which of two equal-sharing principles, equal probabilities of winning a subject-specific money prize, or equal expected money winnings, the information they had in common allowed them to use. (The structures of these environments, expressed in terms of the probabilities subjects bargained over, were always public knowledge; but, in some treatments, subjects were given private information about their money prizes and prevented from communicating it to their partners.)

There were few disagreements when subjects had enough information in common to use only the first principle, but there were significant numbers when they had enough to use either principle. In each case, the frequency of disagreements corresponded closely to a prediction based on the symmetric equilibrium of a coordination game in which bargainers choose independently between the principle or principles they can use; this suggests that the observed disagreements were due to the difficulty of coordination, not to irrationality. These results are especially informative because they were obtained in experimental environments with (except for a general prohibition against self-identification and the prohibition against messages about money prizes when these were not public knowledge) no artificial restrictions on subjects' bargaining strategies. In subsequent experiments, with subjects allowed to communicate only by two stages of simultaneous demands in environments otherwise like those just discussed, it was also found that manipulating subjects' prior expectations about each other's demands exerted a systematic, persistent influence on the agreements they reached.

Even in the experiments reviewed by Jack Ochs and Roth (1989), conducted in environments designed to conform to the assumptions of finite-horizon sequential-offers models, the unique subgame-perfect equilibria in these models predicted poorly. Subjects' behavior appeared to have been heavily influenced by equal-sharing principles, and this made it dangerous to try to extract all of the surplus, as those equilibria prescribe.

Although most of the experiments discussed in Section III used subjects fully informed about the structures of the games they were playing, most of the theoretical models of tacit bargaining that underlie the analyses of explicit bargaining discussed in Section II have incomplete information, typically about bargainers' reservation prices. Some of these have simultaneous-offers structures like the demand game; others have sequential-offers structures. Their central theme is the use of the commitments allowed by the rules of bargaining, together with the risk of disagreement or the cost of delay, to sort bargainers by their reservation prices. Special cases aside, even the models with sequential offers have large sets of equilibria that satisfy standard equilibrium refinements. The usual practice is to study how bargainers learn about each other's private information *within* one of these equilibria. However, strategic uncertainty is unlikely to be less important here than with complete information, and it is unclear to what extent its effects can be captured in an equilibrium selected by applying refinements that are largely insensitive to it. In fact, John Kennan and Robert Wilson (1990) recently surveyed the literature on sequential-offers bargaining under incomplete information with applications in mind, finding significant, systematic differences between its predictions and U.S. and Canadian strike data.

II. Explicit Bargaining

This work builds on models of tacit bargaining like those discussed above, allowing bargainers one or more opportunities to communicate by sending nonbinding messages either before or during the exchange of binding offers and counteroffers. Because a nonbinding message has no direct payoff implications, it can affect outcomes only through the inferences the message receiver draws from the message sender's implied choice among his responses. This allows two roles for such messages in bargaining: signaling bargainers' private information and helping them to coordinate their expectations about how the underlying bargaining game will be played.

The signaling role was first explored, in more abstract settings, by myself and Joel Sobel (1982), and by Jerry Green and Nancy Stokey (1980). These papers showed that there are always uninformative equilibria, in which agents send nonbinding messages unrelated to their private information and these messages are ignored when received, but that there can also exist informative equilibria. A message sent in such an equilibrium means, in effect, "Given my private information, I like what you do when I say this better than anything I could get you to do by saying something else." Normally, equilibria in which the message sender's choice among responses conveys perfect information to the message receiver cannot exist unless their preferences are the same, and only uninformative equilibria can exist when their preferences are too far apart.

Schelling noted that the arguments for Nash's (1953) view that bargaining is primarily a coordination problem remain valid even when cheap talk is allowed, and his analysis of explicit bargaining emphasized its coordination role. He observed that fully informed bargainers have nothing to communicate except what shares of the surplus they expect to receive, and that whatever a bargainer expects, it is to his advantage to convince his partner that he expects more rather than less. It follows that, unlike in pure coordination games, cheap talk cannot coordinate bargainers' expectations *by itself*. Schelling observed that it might still influence bargaining outcomes, either by making it possible for bargainers to use new focal principles to coordinate their expectations, or by allowing them to coordinate their decisions more effectively, thereby expanding their opportunity set.

The coordination role was explored more formally by Joseph Farrell (1987), who studied the Battle of the Sexes game with complete information (which can be viewed as a discrete version of Nash's demand game) preceded by one or more rounds of nonbinding, simultaneous announcements about players' intended strategy choices. He identified a plausible equilibrium in which announcements that would be in equilibrium in the underlying game are accepted as the literal truth, and any other announcements are ignored. (The assumption that announcements are taken literally as long as they are consistent with players' incentives is a powerful equilibrium refinement in cheap-talk models; see, for example, Farrell and Robert Gibbons, 1988, and Roger Myerson, 1989.) In this equilibrium, players' randomized announcements give them a chance of achieving asymmetric coordination in the underlying game, even though they have no way to distinguish between their roles *ex ante*, and are therefore effectively restricted to symmetric strategies in the game taken as a whole. They do better, *ex ante*, than without communication, and still better with additional rounds of communication; but the usefulness of cheap talk is again ultimately limited by the difference between their preferences, and even an unlimited number of rounds does not make coordination certain.

The signaling role of cheap talk has now been studied in bargaining games, by Steven Matthews (1989) in a two-stage sequential-offers model with one-sided private information (see also Farrell and Gibbons, 1988), and by Farrell and Gibbons (1989) and Matthews and Andrew Postlewaite (1989) in a one-stage simultaneous-offers model with two-sided private information, which can be viewed as an incomplete-information generalization of Nash's demand game. The models with two-sided private information allow one or more rounds of simultaneous messages before the underlying game is played; these messages also play a coordination role, in effect allowing bargainers to play different equilibria in the underlying game depending on their reservation prices. Matthews and Postlewaite show that this enormously expands the (already huge) set of equilibria in their model, allowing any outcome that is *ex post* individually rational and incentive compatible.

Finally, Thomas Palfrey and Howard Rosenthal (1990) studied the coordination and signaling roles of cheap talk in n-person games with incomplete information that otherwise closely resemble Nash's demand game, focusing on equilibria like the one studied by Farrell (1987) and obtaining results that generalize his in interesting, sensible ways.

III. Experimental Evidence

As suggested by the discussion above, experiments in which explicit communication was allowed throughout, not just "preplay," are included. Also included are experiments that used coordination games with the perfectly opposed preferences about how to coordinate characteristics of bargaining problems, even some not motivated as bargaining experiments and some conducted with more than two subjects. I have, however, excluded experiments with face-to-face communication. Although some of these have revealed interesting effects, including a suggestion that such communication has some power to bind players and is therefore not truly cheap talk, their results are difficult to interpret because subjects' preferences were evidently partly nonmonetary, hence not fully under experimental control.

The experimental results reported in Roth and J. Keith Murnighan (1982) (also summarized in Roth, 1985) strikingly illustrate the subtle interaction between bargainers' need to find a way to coordinate their expectations, and their strategic use of the information required by focal principles. The experimental environments were like the ones described in Section I that did not artificially restrict subjects' bargaining strategies, but with three significant differences: Some subjects now had an informational advantage over their partners, in that they knew their partners' money prizes as well as their own; the existence of this advantage was sometimes made public knowledge and sometimes not; and subjects were no longer prevented from communicating their private information about prizes to their partners. The results were determined mainly by whether the subject whose money prize was lower knew his partner's prize. When he did not, subjects' agreements closely approximated equal probabilities; but when he did, there was a significant shift toward the probabilities that implied equal expected money payoffs. (The only significant difference between the results for the public knowledge treatments and their non-public knowledge counterparts was a higher frequency of disagreements in the latter treatments when the subject whose money prize was lower also knew his partner's prize.)

These results suggest that most subjects expected to share the gains from their agreements equally according to the most "relevant" principle they could use, given the information they had in common plus whatever additional information they had an incentive to reveal, on the understanding that it would be used to determine the outcome in this way. The transcripts suggest that when only the subject with the higher money prize knew both prizes, and this was public knowledge, the other subject often suspected that he would be treated unfairly. In this case, he knew that his partner could tell whether this was the case, but his partner typically refused to reveal his prize, and the subject then had little choice but to go along with the equal-probabilities agreement that he recognized as his best chance to coordinate, given the information he and his partner had in common.

Russell Cooper et al. (1990) studied the effect of explicit communication in the Battle of the Sexes game, sometimes with one subject allowed to make a single nonbinding announcement about his intended strategy choice before each play of the game, and sometimes with one or three rounds in which subjects were allowed to make simultaneous announcements. When only one subject could make an announcement, he almost always announced the strategy associated with his most preferred equilibrium, and subjects then almost always played the associated strategies. The results when both subjects were allowed to make announcements corresponded fairly closely to the equilibrium in the closely related game studied by Farrell (1987), with subjects doing significantly better than they did with no communication at all and slightly better with three rounds of announcements than with one, but stopping well short of achieving perfect coordination.

Palfrey and Rosenthal also found some support for this view of the coordination process in three-person public goods experiments with one round of simultaneous, nonbinding announcements, based on their incomplete-information generalization of Farrell's 1987 model. They found that subjects had considerable success in efficiently eliciting the contributions required for provision of the public good from the subjects whose privately known costs were lowest. Jeffrey Banks et al. (1988) obtained roughly similar results in experimental environments with similar structures. In each case, significant learning effects were observed in repeated play in tacit versions of the experiments, but not in their explicit counterparts, suggesting that it may be possible to design experiments to measure the effectiveness of cheap talk in reducing strategic uncertainty.

Much of the behavior observed in these experiments appears to have been driven by strategic uncertainty. By contrast, most of the theoretical analyses of explicit bargaining reviewed in Section II either focus on the signaling role of cheap talk, or study the coordination role within an equilibrium selected by refinements that are largely insensitive to strategic uncertainty. The growing experimental literature in this area offers the possibility of developing a theoretical and practical understanding of how strategic uncertainty and cheap talk interact (with or without private information) to determine bargaining outcomes. The theoretical analyses of experimental data in Roth-Murnighan and Roth (1985) show that it may be possible in many cases to model the effects of strategic uncertainty without giving up equilibrium analysis entirely; but, at the very least, this requires explicit attention to bargainers' need to coordinate their expectations, and probably also to the dynamics of the process by which they do so.

REFERENCES

Banks, Jeffrey, Plott, Charles and Porter, David, "An Experimental Analysis of Unanimity in Public Goods Provision Mechanisms," *Review of Economic Studies*, April 1988, *55*, 301–22.

Cooper, Russell et al., "Communication in the Battle of the Sexes Game," *Rand Journal of Economics*, forthcoming 1990, *21*.

Crawford, Vincent P. and Sobel, Joel, "Strategic Information Transmission," *Econometrica*, November 1982, *50*, 1431–51.

Farrell, Joseph, "Cheap Talk, Coordination, and Entry," *Rand Journal of Economics*, Spring 1987, *18*, 34–39.

______ and Gibbons, Robert, "Cheap Talk, Neologisms, and Bargaining," Working Paper No. 500, MIT, July 1988.

______ and ______, "Cheap Talk Can Matter in Bargaining," *Journal of Economic Theory*, June 1989, *48*, 221–37.

Green, Jerry and Stokey, Nancy, "A Two-Person Game of Information Transmission," Harvard Institute of Economic Research Discussion Paper No. 751, March 1980.

Harsanyi, John and Selten, Reinhard, *A General Theory of Equilibrium Selection in Games*, Cambridge: MIT Press, 1988.

Kennan, John and Wilson, Robert, "Strategic Bargaining Models and Interpretation of Strike Data," *Journal of Applied Econometrics*, forthcoming 1990, *5*.

Matthews, Steven, "Veto Threats: Rhetoric in a Bargaining Game," *Quarterly Journal of Economics*, May 1989, *104*, 347–69.

______ **and Postlewaite, Andrew,** "Pre-Play Communication in Two-Person Sealed-Bid Double Auctions," *Journal of Economic Theory*, June 1989, *48*, 238–63.

Myerson, Roger, "Credible Negotiation Statements and Coherent Plans," *Journal of Economic Theory*, June 1989, *48*, 264–303.

Nash, John, "The Bargaining Problem," *Econometrica*, April 1950, *18*, 155–62.

______, "Two-Person Cooperative Games," *Econometrica*, January 1953, *21*, 128–40.

Ochs, Jack and Roth, Alvin, "An Experimental Study of Sequential Bargaining," *American Economic Review*, June 1989, *79*, 355–84.

Palfrey, Thomas and Rosenthal, Howard, "Testing for Effects of Cheap Talk in a Public Goods Game with Private Information," *Games and Economic Behavior*, forthcoming, 1990, *2*.

Roth, Alvin, "Toward a Focal-Point Theory of Bargaining," in his *Game-Theoretic Models of Bargaining*, New York: Cambridge University Press, 1985.

______ **and Murnighan, J. Keith,** "The Role of Information in Bargaining: An Experimental Study," *Econometrica*, September 1982, *50*, 1123–42.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, *50*, 207–11.

Schelling, Thomas, *The Strategy of Conflict*, Cambridge: Harvard University Press, 1960; 2nd ed., 1980.

Bad and Good News About The Sealed-Bid Mechanism: Some Experimental Results

By ANDREW SCHOTTER*

This paper surveys a set of experiments conducted by Roy Radner and myself (1989) and Peter Linhart, Radner, and myself (1989a; b), all of which were conducted to investigate the properties of the sealed-bid mechanism, a mechanism used to structure bargaining under situations of incomplete information. What we find is both bad news and good news for the mechanism. The bad news is that the behavior of subjects seems quite sensitive to the parameters of the prior distribution of types used and the number of rounds over which the experiment is run. As such prior distributions become more skewed (in a sense to be defined below), and the number of rounds in the experiment are increased to 75, bidder behavior becomes less linear. Such movements have a regular pattern, however, and this phenomenon is remarkably consistent.

The good news is that such behavioral shifts do not seem to interfere with the efficiency of the mechanism. This is of course important for the acceptance of the mechanism in the real world, since if efficiency were to change dramatically as we change the parameters of the environment, adoption would have to proceed on a case-by-case basis. The fact that our results imply robustness of efficiencies indicates that, at least on a practical level, the sealed-bid mechanism may be a viable way to structure bargaining in many large-scale economic organizations.

I. The Sealed-Bid Mechanism

To quickly summarize the workings of the mechanism, assume that a potential buyer B and a potential seller S are bargaining over the terms of a possible trade of a single object. If the object is traded, the value to B is V and the cost to S is C. (The seller incurs no cost if there is no trade.) The *sealed-bid mechanism* works as follows: B and S simultaneously choose bids, v and c, respectively. If $v \geq c$, then the trade takes place, and B pays S the price $P = (v + c)/2$, that is, the average of the two bids. If $v < c$, then no trade takes place and B pays S nothing.

Suppose that, at the time of bidding, B knows V but not C, and S knows C but not V. The situation is modeled by supposing that V and C are random variables with a joint probability distribution called the *prior*, that is known to both parties. For all experiments surveyed here, it was assumed that the values and costs of the buyers and sellers are drawn independently from the closed interval [0,100] using the following distributions:

$$P(V) = 1 - ([(100 - V)/100])^{r_1};$$

$$P(C) = (C/100)^{r_2}.$$

By varying r_1 and r_2 from 0 to 1, we can move these distributions from the perfect certainty case ($r_1 = 0$) to the case of a uniform distribution ($r_i = 1$). When $r_1 = r_2$, we call the mechanism "symmetric," while when $r_1 \neq r_2$, the mechanism is "asymmetric." Before the bidding takes place, B observes V but not C, and S observes C but not V. Buyer B's strategy is a function β that determines his bid v for each value of V, and S's strategy is a function ψ that determines his bid c for each value of C. Second-best optimal linear strategies were derived by

---

*New York University, New York, NY 10003. Much of this work has been carried out under NSF grant no. SES-871277 given to New York University. The support of the C. V. Starr Center for Applied Economics is also acknowledged. Finally, I thank Vicky Myroni for writing the programs upon which these experiments were run and Ken Rogoza for valuable research assistance.

Kaylan Chatterjee and William Samuelson (1983) for environments in which the priors were uniform, while Roger Myerson and Mark Satterthwaite (1983) investigated the welfare properties of this mechanism in more general environments.

There are other equilibria to this mechanism as Wolfgang Leininger et al. (1989) and Satterthwaite and Steven Williams (1989) have shown. In fact, there is an infinite number of other such equilibria, some of which involve continuous but nonlinear bid functions, and some of which involve discontinuous-step functions characterized by flat segments to their bidding functions (segments of slope zero).

The Experiments Performed. While space does not permit even a cursory description of the experiments performed, suffice it to say that subjects engaged in an exercise that faithfully captured the characteristics of the sealed-bid mechanism and motivated subjects to perform with salient payoffs. Except for Radner's and my paper, all experiments were performed on a networked set of PCs. In that paper, in all experiments except one, subjects had uniform priors induced upon them ($r_1 = r_2 = 1$) and repeated the experiment 15 times with the same partner. In the exceptional experiment, a 40-round horizon was used and nonuniform priors induced in which buyers had distribution functions skewed to high values and sellers had distribution functions skewed to low costs ($r_1 = 0.4, r_2 = 0.4$).

II. Radner-Schotter

What Radner and I found was that there was indeed at least qualitative support for the linear equilibrium bid strategies. (Similar linear bid functions were also found by James Cox et al., 1986, for sealed-bid auctions). The only exception to this result was our 40-round experiment in which nonuniform value and cost distributions were used ($r_1 = 0.4, r_2 = 0.4$). In these experiments, a modified form of step-function equilibrium was observed in which the bid functions of the buyer and sellers consisted of a two-piece piecewise linear function with a zero-sloped flat segment followed by a nonzero slope linear segment for the sellers and a nonzero sloped segment followed by a flat segment for the buyers.

The question left unanswered by Radner's and my paper was what caused this behavior shift. Since step-function-like behavior was *only* seen in an experiment where a nonuniform distribution was used *in conjunction with* a 40-period horizon, it was unclear whether it was the horizon or the change in the distributions that accounted for the change.

These findings lead Linhart, Radner, and I (1989a) (hereafter, L-R-S) to perform some further experiments.

III. L-R-S (1989a)

In an effort to answer this question, L-R-S (1989a, b) performed two separate series of experiments that varied the prior distribution coefficients and the number of rounds of the experiments. The results of the experiments are summarized as follows. First, L-R-S were successful in replicating the step-function-like behavior exhibited in Radner's and my paper. This behavior was observed quite frequently and in very dramatic fashion. It is a robust feature of the mechanism in practice. Further, it appears that a 75-round horizon is a necessary but not sufficient condition for this behavior to emerge. To insure quasi-step-function behavior, one must move the r_1 and r_2 coefficients away from 1 and increase the experiments horizon to 75 rounds (or at least past 30 rounds).

Second, the bid and ask functions of subjects seemed to fall naturally into three distinctive types which, for lack of better terms, we have called linear, broken, and bent. While these terms are defined more precisely in L-R-S (1989a), a subject's bid (ask) function is called "linear" if it has no significant changes in slope over its domain. A subject's bid (ask) function is called "broken" if it exhibits one significant change in slope at some point, and if it has a slope of zero over some substantial portion of its domain. A subject's bid (ask) function is called "bent" if it exhibits one significant change in slope at some point but has no segment whose slope is zero.

The frequency with which these types appear is clearly related to the horizon and coefficients of the experiment being performed. For example, in the 15-round experiments run, only 2 out of 31 buyers were categorized as anything but linear. (Both of these occurrences appeared in the experiments where both the buyer and the seller had coefficients of .4 in their distribution functions.) Three out of 31 sellers were characterized that way. Hence, when the horizon of the experiment is 15 rounds (and the coefficients of the distribution function are symmetric), almost all buyer and seller behavior can be characterized as linear. In contrast, when the experiments last for 75 rounds (while the symmetry of the distributions functions is maintained), the distribution of types becomes much more diverse with 12 buyers being categorized as linear, 8 as bent, and 5 as broken. The distribution for the sellers is similarly dispersed with 9 being categorized as linear, 7 as bent, and 9 as broken. Going to 75 rounds seems necessary for the diversification of behavioral types. However, it is not sufficient.

In terms of efficiencies, the sealed-bid mechanism seems to be remarkably robust in its ability to generate efficiency levels that do not vary as either the parameters of the experiment change, or as the subjects alter their bidding behavior. In general, across all subject-pairs efficiencies, when measured as the fraction of the total amount of first-best gains from trade captured, averaged in the 80s. Efficiencies are also consistent over the horizon of the 75-round experiments exhibiting no trend toward either improvement or deterioration when we compare the first, middle, or last 25 rounds. In addition, there is no consistent pattern of efficiency change either as a function of the behavioral types of our pairs or as a function of the experiments parameters. A two-way ANOVA test performed in L-R-S (1989a) substantiates this point. In addition, a series of Mann-Whitney U-Tests uncovered no effect of time, or learning on efficiency in the sense that holding the bidding behavior of pairs constant, the efficiencies observed by subjects using these behaviors did not vary significantly between the first, second, and third 25 rounds of our 75-round experiment. Similarly, using the same techniques, we saw no improvement in efficiency over time when we hold the prior distribution coefficients constant. These results substantiate the claim that the mechanism appears to be robust not only to the parameters of the environment but also to the manner in which people behave under it given these parameters.

IV. Linhart, Radner and Schotter (L-R-S 1989b)

Experimental Designs. While L-R-S (1989a) did yield some of the answers to the questions posed above, there are still some questions left unanswered. First, L-R-S (1989a) leaves unanswered the question of why the behavior change observed in Radner's and my paper appears in the first place? Further, are subjects able to infer from the information given to them during the experiment what type of bid function their partner is using, and is this inference important for them? Finally, given these observed changes, are the subjects acting optimally, that is, if a subject knew that his opponent was employing a broken (or linear) bid function, would the bid function observed be a best response?

To answer these questions, L-R-S (1989b) employed an experimental design in which live subject sellers engaged in 75-round sealed-bid experiments with preprogrammed computerized buyers (in some cases the computerized subjects were sellers) employing one of two fixed strategies. These functions were the following: Linear: $b=.82V$; Broken: $b=V$ if $V\leq 60$ and $b=60$ if $60\leq V\leq 100$.

In addition to varying the computerized subjects' bid functions, these experiments were run under two information conditions and using two different sets of priors ($r_1=r_2=1$, and $r_1=.2$, $r_2=.4$). In the high-information condition, subjects were told distribution of *bids* that they could expect to observe from their computerized partner's bid function, while in the low-information condition, they were told only the prior distribution of their opponents values (this was the information condition used in the live experiments of Radner and myself, and L-R-S, 1989a). The bid functions used were in some sense representative of those we observed in L-R-S (1989a) while the coefficients used for the prior distributions were the ones which yielded the most dramatic linear and broken behaviors. Clearly, a 75-horizon experiment was also a necessity. Seventeen experiments were performed.

For each computerized bid function, two separate experiments were performed: one each using different prior distributions ($r_1 = .2$, $r_2 = .4$ and $r_1 = 1$, $r_2 = 1$). The order of these presentations was altered as well. Extensive post-experiment surveys were also used. This design does not help in answering the first question posed above, since it only investigates the behavior of subjects in response to their computerized partner's particular linear or broken strategy. It offers no explanation of why *pairs* of subjects change their behavior as the parameters of the environment change.

More precisely, why do subject pairs choose complementary broken strategies when the coefficients in their prior distributions move sufficiently toward zero (given a long enough horizon)? One conjecture is that when it is common knowledge that the buyer's values are sufficiently skewed toward high realizations, and the seller's costs sufficiently skewed toward low realizations (as in our experiments where $r_1 = 0.2$ and $r_2 = 0.4$), subjects treat these situations as the certainty case:[1] $r_1 = 0$, $r_2 = 0$. More precisely, they treat highly skewed cumulative distributions as degenerate distributions with a point mass at 0 for sellers and a point mass at 100 for buyers. In the certainty case, when one chooses a bid, there is no fear that there are no potential gains from trade. In fact, each subject knows that there exist the maximal gains from trade available. The only question becomes one of distributing these gains from trade, so that, over time, we would expect that in a certainty experiment some commonly agreed to price would emerge. The repeated submission of this agreed to bid and ask would take the appearance of a broken-bid function, but, to properly see the analogy, we will have to observe bids as a function of time and not of realized values and costs, since they have a degenerate distribution. Once implicitly agreed to, such a price furnishes a Nash equilibrium.

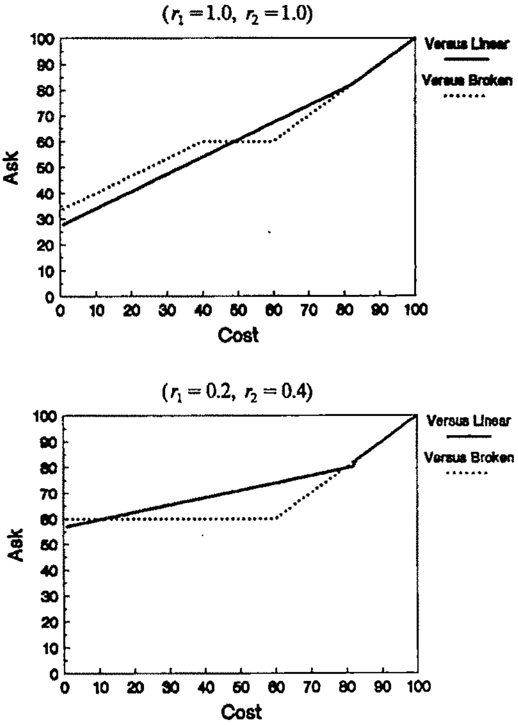
To investigate this conjecture L-R-S (1989b) ran an additional certainty experiment ($r_1 = r_2 = 0$) which we hoped to use as a basis of comparison for our previously skewed and uniform distribution experiments (1989a). In this experiment, it was common knowledge that buyers would have a value of 100 in every round and sellers would have a cost of 0. This was done with two-sided live subjects since we are interested in the effect of certainty on joint subject behavior.

Results. The L-R-S (1989b) experiments in which live subjects play against computerized partners indicate quite clearly that the bidding behavior of live subjects is not influenced to any significant degree by the bidding strategy of their opponents, but rather only by the parameters of the experiment when their partners are programmed to use specific linear or broken bid functions. By this, we mean that if one compares the results of experiments with identical priors but different bidding functions for the computer, the observed bidding functions for subjects do not appear to differ significantly. However, a comparison experiments in which the bidding strategy of the computer is held constant, but the prior distributions change, does indicate a significant amount of change. While this appears strange at first glance, it is not surprising once one investigates the best-response functions of live subjects facing these computerized subjects.[2] For example, Figure 1 presents the best-response

---

[1] For example, when $r_1 = .2$ (.4) there is only a 24 percent (43 percent) chance that a buyer will receive a value less than 75 or a 76 percent (57 percent) chance that he will receive a value greater than 75. Likewise, for $r_2 = .2$ (.4) the seller has only a 24 percent (43 percent) chance of getting a cost greater than 25.

[2] Linhart furnished the calculations upon which this statement is made.



Note: For the best response to the broken strategy, it is assumed that ask = cost for costs greater than 60. Actually, any cost-covering ask is a best response, but asking cost weakly dominates.

FIGURE 1. SELLER BEST RESPONSES VS. LINEAR AND BROKEN BUYER BID FUNCTIONS

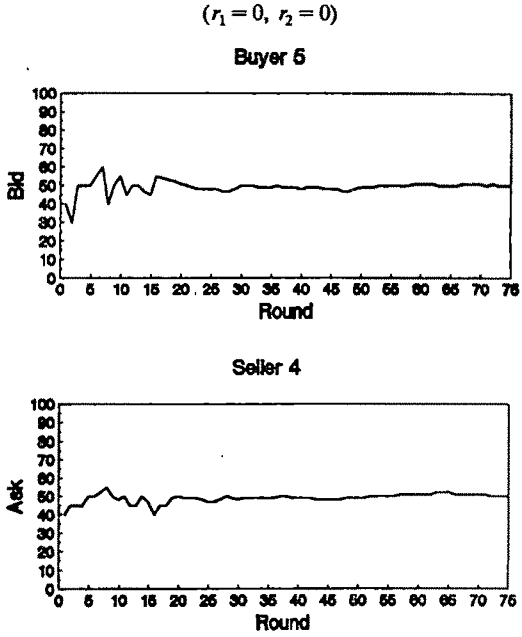


FIGURE 2. BID-ASK FUNCTIONS OVER TIME: THE CERTAINTY CASE

function of our subjects to the strategy of their computerized partners as well as the parameters of the experiment's environment. As we see, bidding behavior is far more responsive to changes in the parameters of prior distributions than changes in the bidding behavior of their opponent. At least, on a qualitative level, it is shown in my paper (1990) that the mean actual behavior of subjects did in fact conform to these best-response functions.

The final conjecture of L-R-S (1989b) concerned the cause for the shift in bidding behavior as the parameters of the experiment changed. It was conjectured that in those experiments where the prior distributions were sufficiently skewed, subjects would behave as they would in the certainty case.

In the certainty case, we expect to see an agreed-to splitting of the gains from trade. Hence, if we plot the bids made by subjects against time, we would expect to see eventually both buyer and seller converging on the same agreed upon price. Figure 2, which presents one example of bid and ask time patterns for one buyer and one seller, substantiates this belief. (Space limitations require that I present only one example. This example is representative of the vast majority of cases, however.)

While in some cases there is a trend exhibited, in general, a constant bid function appears. Equivalent (certainty-like) behavior in the skewed experiments ($r_1 = 0.2$, $r_2 = 0.4$) would generate bid (or ask) functions that, after a certain amount of experimentation, were flat with respect to time, but which exhibited upward spikes (downward spikes) whenever a value (or cost) is realized that is above (below) the agreed upon price. Such behavior would be consistent with our two-piece piecewise linear function projected onto in value bid (cost-ask) space. Figure 3 pre-

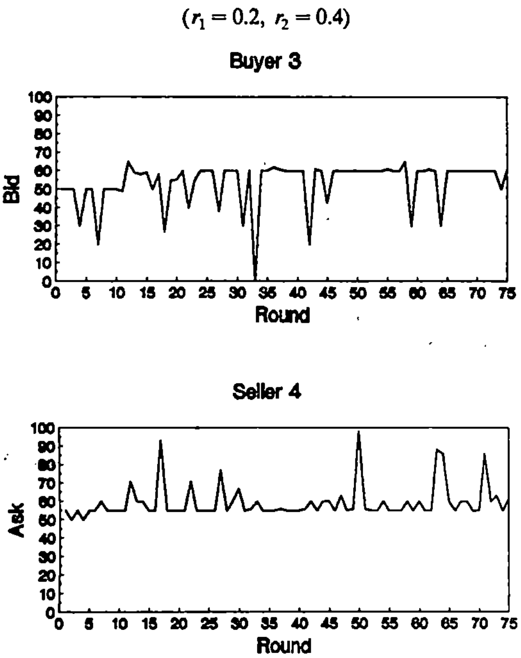


FIGURE 3. BID-ASK FUNCTIONS OVER TIME: THE SKEWED DISTRIBUTION

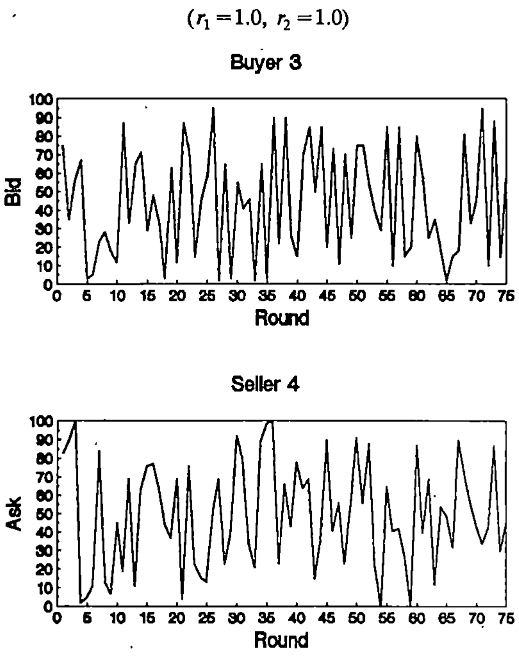


FIGURE 5. BID-ASK FUNCTIONS OVER TIME: THE UNIFORM DISTRIBUTION

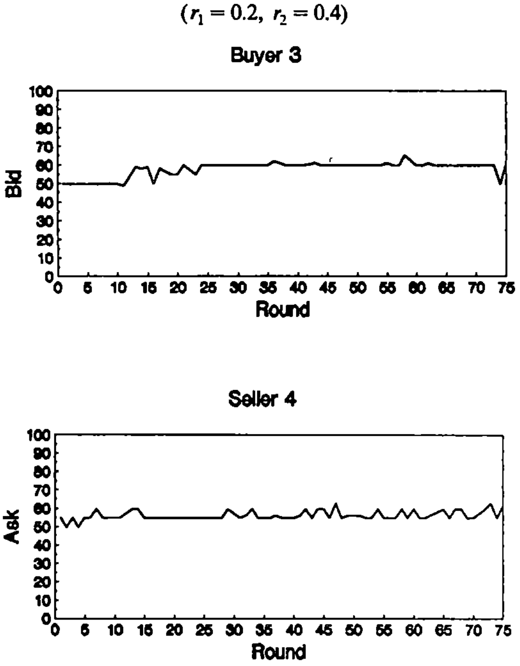


FIGURE 4. BID-ASK FUNCTIONS OVER TIME: THE SKEWED DISTRIBUTION MAJOR SPIKES REMOVED

sents another representative sample of the bid and ask time patterns in the $r_1 = 0.2$, $r_2 = 0.4$ live experiments of L-R-S (1989a). Note that these functions are basically constant with respect to time except for spikes that occur when a value or cost is realized above or below the implicitly agreed to price. These spikes are, of course, necessary in order to avoid losses on the parts of subjects. If we were to eliminate these spikes, we would create Figure 4 which has much of the appearance of the graphs in Figure 2. The comparison of Figures 2 and 4 help substantiate the conjecture that sufficiently skewed distributions may be treated like degenerate certain distributions by subjects.

These diagrams are in sharp contrast to the type of behavior observed in those experiments where all prior distributions are uniform ($r_1 = r_2 = 1$). In these experiments, if subjects bid some linear function of their value or cost, we would expect to see a time graph of bids which was simply a linear transformation of the realizations of these random variables. In Figure 5 we present another representative example of a bid-and-ask pattern from the uniform experiments. Note that this behavior is indeed different from those seen in Figures 2, 3, and 4.

V. Conclusions

What have we learned about the sealed-bid mechanism? First, we know that depending on the environment in which it is used, the behavior of experimental subjects using it may vary dramatically. As the prior distributions used to define costs and values become more skewed, the bidding functions of subjects become more broken. Such changes in bidding behavior seems not to affect the efficiency of the sealed-bid mechanism.

REFERENCES

Chatterjee, Kaylan and Samuelson, William, "Bargaining Under Incomplete Information," *Operations Research*, 1983, *31*, 835–51.

Cox, James, Smith, Vernon and Walker, James, "Theory and Individual Behavior of First Price Auctions," unpublished manuscript, University of Arizona, May 1986.

Leininger, Wolfgang, Linhart, Peter and Radner, Roy, "Equilibria in the Sealed-Bid Mechanism for Bargaining With Incomplete Information," *Journal of Economic Theory*, June 1989, *48*, 63–106.

Linhart, Peter, Radner, Roy and Schotter, Andrew, (1989a) "Behavior and Efficiency in the Sealed-Bid Mechanism: Part 1," mimeo., 1989.

______, ______, and ______, "Behavior and Efficiency in the Sealed-Bid Mechanism: Part 2," in progress, 1989.

Myerson, Roger and Satterthwaite, Mark, "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, April 1983, *29*, 265–81.

Radner, Roy and Schotter, Andrew, "The Sealed-Bid Mechanism: An Experimental Study," *Journal of Economic Theory*, June 1989, *48*, 179–221.

Satterthwaite, Mark and Williams, Steven, "Bilateral Trade With the Sealed-Bid *k*-Double Auction: Existence and Efficiency," *Journal of Economic Theory*, June 1989, *48*, 107–33.

Schotter, Andrew, "Bad and Good News About the Sealed-Bid Mechanism: Some Experimental Results," working paper no. 90-0, C. V. Starr Center for Applied Economics, New York University, January 1990.

The Transition from Bargaining to a Competitive Market

By STEVEN R. WILLIAMS*

How few traders constitute a bargaining problem or, alternatively how many traders constitute a market? Intuitively, people who meet to trade face a bargaining problem if several of them, with opposing interests, can influence the outcome of trade through their behavior. Traders constitute a competitive market only if the effect of any trader on the outcome of trade is insignificant. Analysis of bargaining using noncooperative game theory has been a lively research topic over the past ten years, while the formalization of perfectly competitive markets using a continuum of traders is now several decades old. In this paper I discuss a line of research that studies the transition between these two theories; in particular, a result from Mark Satterthwaite's and my paper (1989b) is presented that shows how a bargaining problem is transformed into a market as the number of traders increases.

Incomplete information is an essential feature of the model that I discuss. In order to explain how traders achieve a competitive equilibrium, it is standard to assume that all potential gains from trade are commonly known at the outset. In contrast, each trader in the model that I describe privately knows his own preferences. A second feature is that a market with any finite number of traders on each side is modeled, rather than one with a continuum of traders. These two features together make strategic behavior seem especially likely. Price-taking behavior is not an axiom here; instead, the objective is to prove that a trader's equilibrium behavior is to increasingly act as a price taker as a market grows in size, and that only a small number of traders is needed on each side to compel each trader to act in this way.

Though the background of this research will be discussed later in more detail, it is basically motivated by the intuition most of us share that a market can work pretty well despite having a relatively small number of traders who strategically act on their private information. While the axioms of perfect competition are, of course, never satisfied in the real world, I believe that this theory provides insight not only into immense markets but also into smaller markets. The research that I describe is another attempt to substantiate this belief.

I. The Model, Trading Procedure, and Main Result

I consider the following elementary trading environment. Traders can exchange money for discrete units of a homogeneous good. For $m \geq 1$, there are m buyers, each of whom wishes to buy at most one unit, and m sellers, each of whom has one item to sell. A trader's preferences are given by his reservation value for an item. Let v_b denote a buyer's reservation value and let v_s denote a seller's. A buyer's utility is v_b minus his payment when he trades, and zero when he does not; similarly, a seller's utility is what he receives minus v_s when he trades, and zero when he does not. Traders are therefore risk neutral (though this is not crucial to the result).

The following procedure is used to organize trade. Each seller submits an offer while each buyer submits a bid. The offers and bids are ordered in a list, $s_1 \leq s_2 \ldots \leq s_{2m}$. For the moment, assume that the offers and bids are distinct. If there are q sellers' offers at or below s_m, then there are also $m-q$ buyers' bids at or below this number; as a consequence, the remaining q bids must be at or above s_{m+1}. A market-clearing price can thus be chosen in the interval $[s_m, s_{m+1}]$. I focus here upon the procedure that selects s_{m+1} as the price; trade then occurs between buyers who bid at least s_{m+1} and sellers whose offers were strictly less than this value. This is called the *Buyer's Bid Double Auction* (BBDA) because in the bilateral ($m=1$) case, the buyer's bid is the price at which trade occurs. Alternative rules for selecting the price are mentioned at the end of this section.

---

*Economics Department, Northwestern University, Evanston, IL 60208. This research was supported by NSF grant no. SES-8705649. I thank Bob Anderson, Stan Reiter, and Mark Satterthwaite for helpful comments.

Ties in the list of offers/bids are a rather technical point. Satterthwaite and I (1989b) show that ties may cause shortages, that is, the demand at price s_{m+1} may exceed supply. The definition of the BBDA is completed by assuming that when shortages occur, the available supply is distributed among the buyers with the size of the buyer's bid determining his priority. If necessary, the allocation is completed by using a fair lottery to distribute the remaining supply among buyers who submitted the same bid.

The technicalities of ties should not obscure the key point: trade occurs in the BBDA at a market-clearing price that is computed using revealed supply and demand (as expressed by the offers and bids). This is the classic analysis of trade applied to an elementary trading environment. Aside from the unitary supply/demand of each trader, the only way that this differs from a standard model of an undergraduate microeconomics class is that price is computed here using revealed information. It is not assumed that traders tell the truth, or that the true supply and demand curves are publicly known.

In order to study the impact of private information upon trade, behavior in the BBDA is modeled using the theory of Bayesian games. Each buyer's (seller's) reservation value is independently drawn from a distribution $F_b(F_s)$ on $[0,1]$ that is continuously differentiable with nonzero density on this interval. These distributions are common knowledge. A trader's strategy is a rule that specifies his offer/bid as a function of his reservation value. A seller in the BBDA trades only if his offer is less than the market price; it is easy to show that, because a seller cannot influence the price at which he trades, his unique dominant strategy is to honestly set his offer equal to his reservation value. Assuming that each seller adopts this strategy, the result that I describe concerns Bayesian-Nash equilibria in which each buyer uses the same strategy B (i.e., given that each seller acts honestly and every other buyer uses the strategy B, a buyer with reservation value v_b maximizes his expected utility by bidding $B(v_b)$).

While a seller chooses to truthfully set his offer equal to his reservation value, it is easy to show that a buyer has an incentive to underbid; in equilibrium, $B(v_b)$ is less than v_b for positive values of v_b. Even though a mutually profitable deal exists between a buyer and a seller, they may fail to trade because the buyer's bid can be less than the seller's offer. The unfortunate effect of strategic misrepresentation by buyers is inefficiency: all potential gains from trade may not be achieved.

The main result that I discuss suggests that the loss due to strategic misrepresentation is significant only when the size of the market (as given by the number m on each side) is very small. Satterthwaite and I (1989b) show that in an equilibrium of the form described above the amount of misrepresentation by a buyer is $O(1/m)$. Formally, there exists a function $\kappa(v_b;F_b,F_s)$ that is continuous in v_b such that, in any equilibrium strategy $B(v_b)$ in the market of size m,

$$v_b - B(v_b) \leq (\kappa(v_b;F_b,F_s))/m$$

at every v_b in $[0,1)$. This bound can be used to show that the percentage of the potential gains from trade that are not achieved is $O(1/m^2)$. It is worth emphasizing that these rates of convergence hold for any equilibrium of the specified form.

These rates suggest that the BBDA can be almost fully efficient even when relatively few traders are present. This is supported by the following example. Suppose that each trader's reservation value is independently drawn from the uniform distribution on $[0,1]$. It can be shown that $B(v_b)=mv_b/(m+1)$ defines an equilibrium in the market with m traders on each side. I now compare the efficiency of trade under this equilibrium with trade under another set of rules. The *fixed-price mechanism* begins by fixing a price. Each trader indicates whether or not

TABLE 1—THE RELATIVE INEFFICIENCY OF THE FIXED-PRICE MECHANISM AND THE BBDA

Mechanism	$m=2$	$m=4$	$m=8$	$m=10$
Fixed-Price ($p=.5$)	21.87	18.26	14.62	13.50
BBDA	7.42	1.75	0.42	0.26

Notes: For markets of different sizes, the table gives the percentage of the total expected gains from trade that are not achieved in these two mechanisms when reservation values are drawn from the uniform distribution on $[0,1]$. The figures for the BBDA assume that a buyer's strategy in the market with m traders on each side is $B(v_b)=mv_b/(m+1)$.

he is willing to trade at this price; of the available supply and demand, as many trades as possible are then made at this price by randomly matching the buyers and sellers who have expressed their willingness to trade.

No priority is given to the magnitude of one's offer/bid, which insures that each trader's dominant strategy is to truthfully reveal how much his reservation value is above or below the chosen price. The perverse effect of this matching rule is inefficiency, for achieving all potential gains from trade requires a mechanism to assign priority in trade to buyers with high values and sellers with low values. The price that ought to be chosen in the uniform case is $p=.5$ because this is what the price would be if there were a continuum of traders on each side of the market with uniformly distributed reservation values. While the BBDA and the fixed-price mechanism produce the same efficient outcome in this continuum market, they are easily distinguishable in small markets because the outcome of trade in the fixed-price mechanism converges so slowly to efficiency.

Table 1 (drawn from Table 6.1 of Satterthwaite's and my 1989b paper) compares the percentage of the potential gains from trade that are inefficiently not achieved under the two procedures in markets of different sizes. Notice that the BBDA is almost fully efficient with as few as eight traders on a side, while the fixed-price mechanism fails to achieve almost 15 percent of the potential gains from trade in this market. The pattern in this example is summarized by a comparison of the rates of convergence of these two procedures, for the relative loss in the fixed-price mechanism is $O(1/m^{1/2})$ rather than $O(1/m^2)$ as in the BBDA.

Two generalizations of the BBDA convergence result should be noted. First, it is extended in my earlier paper (1988) to markets in which there are m buyers and n sellers. Strategic misrepresentation is shown to be $O(1/\min(m,n))$ in such a market. Fix m,n and consider the market with tm buyers and tn sellers for $t\geq 1$. This bound implies that the relative loss in this replication is $O(1/t^2)$. Second, the bound on misrepresentation given above holds in the case of risk-averse buyers who share a common utility function. This is suggested by the following intuitive argument. In lowering his bid below his reservation value, a buyer weighs the marginal expected gain from lowering the price that he pays against the marginal expected loss from missing out on a profitable trade. A risk-averse buyer attaches relatively more weight to this potential loss than a risk-neutral buyer, which compels him to misrepresent less than the risk-neutral buyer.

Several qualifications on the BBDA convergence result should be mentioned. The assumption that players share common knowledge beliefs is restrictive in almost any Bayesian model, but (as Robert Wilson, 1987, emphasized) it seems especially so when the players are traders in a market. The assumption that each buyer uses the same strategy may or may not be a strong one, for to my knowledge nothing has been proven about equilibria in which buyers use different strategies. Finally, the independence of the reservation values is also problematic. Private information makes the outcome of trade completely inefficient in George Akerlof's (1970) famous example of a market for lemons. This is proven without assuming that the traders act strategically and with no restriction on the number of traders. Satterthwaite's and my analysis of the BBDA and Akerlof's analysis are similar in that both apply the classic notion of trade at a price that equates supply with demand to a market in which traders have incomplete information. Akerlof's example does rest upon the extreme assumption that each potential seller has private information about his item that completely determines its value to any other trader. The lemons principle, however, is so useful and instructive that it cannot be so easily dismissed. The question of how the kind of private information affects the outcome of trade is clearly worth investigating.

I conclude this discussion of the BBDA convergence result by mentioning a possible extension of this result to a broader family of procedures. Recall from the definition of the BBDA that a market-clearing price can be chosen in the interval $[s_m, s_{m+1}]$ between the mth and the $(m+1)$st largest offer/bids. For $k \in [0,1]$, call the procedure the selects $(1-k)s_m + ks_{m+1}$ as the price the *k-double auction*. The BBDA is the $k=1$ double auction, and of course all of its properties have direct analogues for the $k=0$ double auction. The more interesting case is when $k \in (0,1)$, for a trader on either side of the market may then influence the price at which he trades. Traders on both sides therefore strategically misrepresent their reservation values. While this is more interesting as a model, it is also considerably more difficult. The bilateral case $(m=1)$ has already been thoroughly investigated; Wolfgang Leininger et al. (1989) and Satterthwaite and I (1989a) have shown that there exists in this case a great variety of equilibria of vastly varying efficiency. There are few results on the multilateral case. Computational evidence does suggest that misrepresentation converges to zero in any sequence of equilibria in each of which (*i*) traders on each side of the market share a common smooth strategy, and (*ii*) trade occurs with positive probability. It thus seems that the basic theme of the BBDA convergence result generalizes despite the multiplicity of equilibria in these double auctions.

II. Some Comparisons

The BBDA convergence result epitomizes a line of research whose goal is to evaluate the performance of realistic procedures for organizing trade when traders act noncooperatively and have incomplete information. In this section I examine this statement with the focus upon the words "realism," "noncooperative," and "incomplete information."

I first consider the issue of realism from a practical viewpoint: do k-double auctions resemble actual procedures for organizing trade? One must distinguish the trading environment from the procedure to answer this question. Consider trade in a more general environment in which a trader may want to sell or buy multiple units. Each trader submits an offer/bid and a quantity that he is willing to sell/buy; supply and demand curves are then constructed from which a market-clearing price can be selected. This procedure is commonly used to organize trade when a dynamic, ongoing market is inappropriate or unnecessary. The Soviet government, for instance, recently ran such a market for the exchange of dollars and rubles among Soviet firms in order to learn about the true dollar/ruble exchange rate. Assuming that the rule for picking the price is linear, then this procedure is exactly a k-double auction when each trader wants to sell or buy a single unit. My point is that k-double auctions do resemble real procedures; it is the assumption of unitary supply/demand that is so restrictive. Of course, it would be interesting to study markets in which multiple units or even a variety of goods can be traded. It would be even more interesting, I believe, to study models of markets in which time plays an essential role, for many real-world markets (such as stock markets) are essentially dynamic. These tasks, however, are formidable; k-double auctions are a reasonable place to begin a study of strategic behavior in small markets.

I now consider realism from a different viewpoint—that of a theorist trying to nudge game-theoretic models of trade towards reality without making them intractable. A common flaw of such models is that the rules for trading are defined in terms of either the solution concept or the environment. For example, inefficiency in the fixed-price mechanism converges to zero only if the price is chosen properly, and choosing the proper price requires some knowledge of the distributions from which traders' reservation values are drawn. Revelation mechanisms share this flaw, for they are essentially defined in terms of the Bayesian model. In contrast, the rules of a k-double auction are independent of the Bayesian-Nash solution concept. The Bayesian model is imposed only after the rules are completed, as one way to predict the outcome of trade. As Wilson emphasized, the rules of real markets are not defined in terms of features of the environment such as traders' beliefs; instead, they are durable, seemingly flexible enough to work well even as the environment changes. A task for theory is to explain how simple sets of rules may permit reasonably good performance for a variety of environments. In this light, the simplicity of k-double auctions is a virtue. The BBDA convergence rate clearly contributes to this task in that it holds for a large class of distributions.

To conclude the paper, I focus upon the words "noncooperative" and "incomplete information." To emphasize their significance, I compare the BBDA convergence result with the well-known results of Gerard Debreu (1975), Birgit Grodal (1975), and others on the rate of convergence of the core to the competitive equilibria. Three points of comparison between these results should be emphasized. First, the core presumes that traders cooperate to achieve all gains from trade regardless of the number of traders that are present. This is typically supported by assuming that traders have complete information. In contrast, the BBDA model is rooted in a result of Roger Myerson and Satterthwaite (1983) that basically asserts that no trading procedure can be both fully efficient and individually rational when traders have incomplete information and act noncooperatively. Second, how the core is achieved by traders is not modeled, while the BBDA is an explicit procedure for trading. To summarize the first two points, the core convergence results assume away the difficulty of achieving efficiency, which is the focus of the BBDA result. The last point concerns the greater generality of the trading environment in the core convergence results. Such results are proven using an Arrow-Debreu environment, while (as stressed above) the environment in the BBDA result is very special. At present, this is the price that must be paid for the added features of the BBDA model.

REFERENCES

Akerlof, George, "The Market for Lemons: Qualitative Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, *84*, 488–500.

Debreu, Gerard, "The Rate of Convergence of the Core of an Economy," *Journal of Mathematical Economics*, March 1975, *2*, 1–8.

Grodal, Birgit, "The Rate of Convergence of the Core for a Purely Competitive Sequence of Economies," *Journal of Mathematical Economics*, June-September, 1975, *2*, 171–86.

Leininger, Wolfgang, Linhart, Peter and Radner, Roy, "Equilibria of the Sealed-Bid Mechanism for Bargaining with Incomplete Information," *Journal of Economic Theory*, June 1989, *48*, 63–106.

Myerson, Roger and Satterthwaite, Mark, "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, April 1983, *29*, 265–81.

Satterthwaite, Mark and Williams, Steven, (1989a) "Bilateral Trade with the Sealed-Bid k-Double Auction: Existence and Efficiency," *Journal of Economic Theory*, June 1989, *48*, 107–33.

______ **and** ______, (1989b) "The Rate of Convergence to Efficiency in the Buyer's Bid Double Auction as the Market Becomes Large," *Review of Economic Studies*, October 1989, *56*, 477–98.

Williams, Steven, "The Nature of Equilibria in the Buyer's Bid Double Auction," Northwestern University CMSEMS DP No. 793, 1988.

Wilson, Robert, "Game Theoretic Analyses of Trading Processes," in Truman Bewley, ed., *Advances in Economic Theory, Fifth World Congress*, Cambridge: Cambridge University Press, 1987.

THE NATIONAL RESEARCH COUNCIL'S REPORT ON THE STATUS OF BLACK AMERICANS, 1940–85[†]

A Common Destiny: How Does It Compare to the Classic Studies of Black Life in America?

By THOMAS D. BOSTON*

> We cannot close this description of what a study of a Negro community should be without calling attention to the study which best meets our requirement, a study which is now all but forgotten. We refer to W. E. B. Du Bois, *The Philadelphia Negro*, published in 1899.
>
> [Gunnar Myrdal, 1944, p. 1132]

A classic has been called a book that is "often referred to but seldom read." If this is so, then Myrdal's *An American Dilemma* (1944) and Du Bois' *Philadelphia Negro* (1899, 1967) are excellent examples. While most economists consider Myrdal's book the authoritative study of black life in America, few have even heard of Du Bois' inquiry. This discussion is designed to accomplish two objectives. The first is to compare the National Research Council's (NRC) report on the status of black Americans (*A Common Destiny*, Gerald Jaynes and Robin Williams, 1989) to earlier classic studies. Second, to suggest that economists would benefit greatly by broadening their frame of reference to include some less well-known but equally seminal investigations of black life.

The controversy surrounding the commissioning of the NRC report is reminiscent of the criticism encountered by Du Bois one century earlier when he undertook his monumental Philadelphia study. Reflecting upon that experience, Du Bois notes that whites said "Why study the obvious?," while blacks said "Are we animals to be dissected, and by an unknown Negro at that?" (1968, p. 197). Yet his examination of the Philadelphia Negro is so thorough that it has withstood a century of criticism. Today this inquiry is considered to be the forerunner of all studies of black life. "The problem lay before me," Du Bois observed, "Study it. I studied it personally and not by proxy. I sent out no canvassers. I went myself. Personally I visited and talked with 5,000 persons" (1968, p. 198).

Does the NRC report measure up to the Philadelphia study? Or to others Du Bois wrote such as *Black Reconstruction in America* (1935, 1964) or to the fifteen volumes he edited between 1897 and 1915 which constitute the *Proceedings* of the Annual Conferences on the Negro Problem sponsored and published by Atlanta University? How does it compare to Charles Johnson's investigations of black life (1930, 1934, 1943, and with E. Embree and W. Alexander, 1935)? Is it truly an extension of Myrdal's *An American Dilemma*?

Because the NRC assembled a nationally prominent team of researchers to conduct a collaborative inquiry into the status of black Americans, it is inescapable that the results will be compared with earlier collaborative efforts. For the editors, Jaynes and Williams, this is both a challenge and a burden. After all, while the influence of Myrdal's study has pervaded the social sciences, criticisms of it

---

[†] *Discussants*: James Stewart, Pennsylvania State University; Gerald Jaynes, Yale University.

*Associate Professor of Economics, College of Management, Georgia Institute of Technology, Atlanta, GA 30332.

continue even today (D. Southern, 1987, pp. 261–92).[1]

The Myrdal study has been much criticized for its idealist perspective and for not including black researchers among its directors. Even before the NRC report was published, critics drew parallels between its shortcomings and the Myrdal study (see B. Chachere et al., 1987; T. Duster, 1987). The perspective and composition of a collaborative research team are important because they determine which questions are asked and who does the asking. Therefore, dissimilarities in ethnicity, social standing, and political empowerment between the group being studied and the group conducting the study are obvious sources of tension (see Duster). Critics felt that the opposition to affirmative action and government activism of some members of the NRC's Committee on the Status of Black Americans would bias the study's result.

Whether these early criticisms forestalled such conclusions is unclear. But the findings of the report do not support an antigovernment, culture of poverty paradigm. It concludes that low earnings rather than changes in the family structure are primarily responsible for the increasing incidence of poverty among blacks since the 1970s (see Jaynes-Williams, p. 281). Further, its opening section affirms that government support programs are not mainly responsible for the extent of female-headed households, high birth rates among unmarried women, or low labor force participation among black males. Instead it argues that discrimination, the absence of opportunity, as well as a stagnant economy play pivotal roles in these developments (Jaynes-Williams, p. 10).

Comparisons between the NRC report and earlier studies can be made by considering the organization of their research teams, their methodologies, their empirical findings and/or their interpretations, conclusions, and policy recommendations.

My primary objective is to compare methodologies and I use the ideas of Du Bois as my reference point. Though often ignored, Du Bois' contributions to studies of black life are perhaps the most significant of this century. In an introduction to the 1967 edition of *The Philadelphia Negro*, Digby Baltzell asserts that the book's emphasis on class and social environment as major causal agents in personality formation established a new tradition in American sociology. Additionally, he notes that Du Bois influenced the origins, methods, and theoretical perspectives of major studies conducted by the Robert Park School of urban sociology at Chicago and the W. Lloyd Warner School of community studies at Harvard and Chicago (see pp. *xxv* and *xxvi*).

I. Methodological Framework of Du Bois

In his *Autobiography*, Du Bois outlines a methodology for studies of black life (1968: see chs. 12, 13, and 18). What he objects to most are static approaches that simply list a "vague mass of so-called Negro problems" (p. 217). The fundamentals of his methodological framework are:

1) The study should be divided into two essential parts with the first part examining the social group and the second part investigating the group's social environment (including its economic and political relations).

2) The conditions of the group should be revealed as symptoms rather than as causes. Furthermore, its striving and long historic development should be outlined as opposed to simply conceptualizing it as a "transient," "inert," or "sick body of crime."

3) Blacks must be understood as having the same endowments and capacity for culture and self-help as other ethnic groups.

4) Society must be comprehended as an everchanging developing entity, rather than as a fixed social structure. The part played by social relations in personality formation and the factors associated with socioeconomic class membership must be emphasized.

5) The economic aspects of world movements must be considered and the race prob-

---

[1] Since most readers are familiar with the Myrdal study, my discussion will center mainly on less familiar but equally seminal studies of black life.

lem linked with global and national developments.

6) The study should be a scientific investigation into social conditions, and serve as a basis of general information and further study rather than as an organ of social reform; although reform movements should be based on the study.

7) Black American researchers must be at the center of the research team.[2] Furthermore, the study must be adequately funded and directed by a national planning institute whose job it is to gather, interpret, and publish the facts.

8) The study should be divided into ten topical areas with a different topic reported each year until the entire cycle has been completed. The full cycle should then recur continually for an entire century, but each time on a broader, more exact basis. Individual topics should be examined simultaneously but reported every ten years. The purpose is to derive a dynamically moving picture of black life rather than an occasional snapshot.

The Philadelphia Negro, the Atlanta University studies, and *Black Reconstruction* are excellent applications of this methodology. For example, the latter study starts with a detailed examination of all classes and class relations in the slave system, it next considers the peculiarities and contradictions of slavery as a system of production, and then examines the economic, political, and social conflicts that gave rise to the Civil War. Unlike Myrdal, Du Bois does not see reconstruction as the fulfillment of an "American Creed" wherein individuals sought to steer a nation gone astray back to the moral principles embodied in the Constitution. Instead, after identifying all of the conflicting interests, he demonstrates that a primary force behind reconstruction was the desire by northern finance and industry to establish southern governments amenable to their economic domination (see especially ch. 14). This strategy could be accomplished best by subduing the slave oligarchy through radical reconstruction governments and empowering freed slaves. The equality of blacks created thereby is seen as the outcome of a political alliance between northern industry, freed slaves, and the federal government in order to suppress the counterrevolution of the defeated slave oligarchy.

II. The Johnson Studies

Johnson's seminal study of black life, *The Negro in American Civilization*, was published fourteen years before the Myrdal study. Additionally, he published a second study (1943) that was used as memoranda for Myrdal's *American Dilemma*. The methodology and structure of Johnson's studies (1930, 1934, 1943) are very similar to the NRC report, that is, they are largely descriptive compendiums of factual data, and quite unlike the studies of Du Bois. However, Johnson et al.'s short and very insightful analysis, *Collapse of Cotton Tenancy*, employs a typical Du Boisian methodology.

This latter study describes each strata of the land tenant system and outlines the conflicts in economic interest that gave rise to the demise of cotton supremacy in the black belt. It also discusses the social and cultural web of racial domination and the Jim Crow laws that were employed to reinforce and preserve the system.

III. *A Common Destiny*: How Does It Compare?

This study opens with a Myrdalian thesis and antithesis. For Myrdal, the dilemma in American society is the divergence between its constitutional creed and the reality of black life. In *A Common Destiny*, a "new dilemma" is identified. Emerging after the civil rights movement, it is considered to be the contradiction between the aspirations of black Americans for equal opportunity and the actual attainment of it in all sectors of life. Beneath this new dilemma, there still exists a divergence between the social principles espoused by whites and their individual practices which deny equality of treatment to blacks (Jaynes-Williams, pp. 4, 5).

---

[2] Du Bois thought it illogical that by 1944 the center of gravity of research on black life had shifted away from African-Americans and their institutions (1968, p. 313).

The present state of black-white relations is explained by the mutually reinforcing history of racial inequality and the disadvantaged conditions under which many blacks live. Equally important is political activism among blacks, resistance to social change on the part of whites, and changes in the state of the economy. Regarding the latter, the 1940s through the 1970s are considered to be the most significant periods of racial advance because of the tremendous economic growth that occurred. Alternately the slowdown in the economy in the post-1973 period caused a stagnation in opportunities and advancement among blacks which is particularly detrimental to the lowest-income and least-skilled individuals. The most important interdependent events shaping the relative economic status of blacks are: 1) urbanization and northern migration between 1940 and 1970; 2) the civil rights movement; 3) unprecedented economic growth between 1940 and the early 1970s, 4) finally, the significant slowdown in economic growth since the 1970s.

The NRC report affirms that barriers to full opportunity continue to exist in the form of discrimination, residential segregation, and exclusion from social networks that facilitate economic and social mobility. At the same time, it rejects conservative views on the influence of government-support programs, and focuses instead on the economic and political relations that deny opportunities to blacks. Most problems are interpreted as derivatives of discrimination and economic stagnation rather than of culture.

The NRC report does not identify the dynamic changes in the United States and world economies nor their implications for the relative status of African-Americans. It is one thing to recognize that the American economy is stagnating. It is quite another to uncover the dynamics of this stagnation and the particular place black labor occupies within these changes.

The NRC report devotes very little attention to the influence of de jure and de facto discrimination on the historical development of the black capitalist class. Yet, this is one place where racial barriers have been most devastating. Another conspicuous omission is the influence of the Black Power movement on racial identity and social change. As do most studies, the NRC report has excluded the Black Power movement from the recent history of African-Americans in favor of the more comfortable civil rights movement.

Finally, the role of social class is mentioned frequently, but the economic and social dynamics of class relations are not elucidated. Class analysis in the NRC report is reduced to a simple discussion of income boundaries and occupational status (see pp. 169; 274–77). Interested readers are invited to compare this discussion to the very lucid description of classes by Du Bois (1899, 1967; pp. 309–21).

Despite these criticisms, *A Common Destiny* is a fine empirical account of the contemporary status of African Americans. Policy recommendations were not a part of its mandate and it makes very few suggestions explicitly. Its major contribution is that it provides a comprehensive reference for much of the latest orthodox research on black Americans. Even though it failed to consider radical literature, it did not provide fuel for the conservative assault on black social advancement. It falls short of uncovering the underlying social forces that are essential to Du Bois' methodology, but it is not entrapped by the idealism of Myrdal's study. Despite what it could have been or what our individual interests would have preferred it to be, in the final analysis it has achieved exactly what it originally set out to accomplish. It is a summary and an interpretation of a large body of data on the position of blacks in American society since the eve of World War II. I wish that it were more, but I am also quite satisfied with what it is.

REFERENCES

Chachere, B., Smith, R. and Walters, R., "Causes for Alarm," *Society*, January/February 1987, *24*, 22–28.

Du Bois, W. E. B., *Black Reconstruction in America: 1860–1880*, Cleveland: Meridian Books, 1935/1964.

______, *The Philadelphia Negro: A Social*

Study, New York: Shocken Books, 1899/1967.

______, *The Autobiography of W. E. B. Du Bois*, New York: International, 1968.

Duster, T., "Purpose and Bias," *Society*, January/February 1987, *24*, 8–12.

Jaynes, G. and Williams, R., *A Common Destiny: Blacks and American Society*, Washington: National Academy Press, 1989.

Johnson, C. S., *The Negro in American Civilization: A Study of the Negro Life and Race Relations in the Light of Social Research*, New York: Henry Holt, 1930.

______, *Shadow of the Plantation*, Chicago: University of Chicago Press, 1934.

______, *Background to Patterns of Negro Segregation*, New York: Thomas Y. Crowell, 1943.

______, **Embree, E. and Alexander, W.**, *Collapse of Cotton Tenancy*, Chapel Hill: University of North Carolina Press, 1935.

Myrdal, G., *An American Dilemma: The Negro Problems and Modern Democracy*, New York: Harper & Row, 1944.

Southern, D., *Gunner Myrdal and Black-White Relations: The Use and Abuse of an American Dilemma, 1944–1969*, Baton Rouge: Louisiana State University Press, 1987.

Blacks, Hispanics, and White Ethnic Groups: Are Blacks Uniquely Disadvantaged?

By REYNOLDS FARLEY*

The National Academy's *A Common Destiny* described black Americans, and observed that while substantial progress was made, blacks remained far behind whites on social, health, and economic indicators.

Are blacks unique? If we consider Hispanics, other racial minorities, and white ethnic groups, will we find that some groups are even less prosperous than African-Americans? Is the black-white comparison too simple? This paper, part of my larger investigation of racial-ethnic differences (1989), addresses these issues.

I. Defining Racial and Ethnic Groups

The 1980 Census asked persons to identify their race by checking one of 15 circles. All respondents also answered a question about their Hispanic origin. A 19.3 percent sample answered a question about their ancestry. This query, used for the first time in 1980, replaced questions about birthplace of parents. Respondents could write any terms they wished, and typically the first two were coded. There was no allocation for nonresponse to the ancestry question, and about 11 percent left the question unanswered. Another 6 percent wrote the single term, American.

An exhaustive and mutually exclusive array of racial/ethnic groups was created using these questions (see also Stanley Lieberson and Mary Waters, 1988, ch. 2). First, 4 Hispanic groups were defined from the Spanish-origin question. Next, non-Hispanics were classified by their race. Aleuts and Eskimos were grouped with American Indians (by race) to form an American Natives category, and then 7 nonwhite racial groups were identified. A residual category of "other races" included Hawaiians, Guamanians, and other infrequently reported races.

Those who identified themselves as white by race and were not Hispanics were classified by their first reported ancestry. I identified 37 white ethnic groups, each having 60,000 persons between the ages 25–54; a number that allowed the fitting of models of earnings using the one-percent sample. The list of white ethnic groups includes American Indians. There were numerous individuals (approximately 2 million) who selected white as their race and then wrote American Indian for their ancestry. This group should be distinguished from the smaller Native American group (C. Matthew Snipp, 1980, ch. 2). The list includes a category of whites who wrote American for their ancestry, those who left the ancestry question blank, and whites not elsewhere classified (NEC): people whose first ancestry was not among the 36 used in this analysis. English and German each made up about 15 percent of the population. The third largest group, blacks, was the only other one comprising 10 percent or more. The 4 smallest groups (Croatian, Rumanian, Lebanese, and Armenian) each comprised less than .1 percent.

II. Economic Characteristics

Table 1 shows the percent in households whose pre-tax cash incomes either fell below the poverty line or exceeded six times the poverty line (a group I label affluent). To be impoverished, a household of four had to have an income of less than $11,600 in 1987 dollars; to be affluent, in excess of $69,600.

The white ethnic groups (except Americans and American Indians) had poverty rates between 4 and 8 percent, or well below the national average of 11 percent. Among Filipinos, Japanese, Koreans, and Asian Indians, poverty rates were equivalent to those of white ethnic groups, but 6 groups were distinguished by elevated poverty. The most

---

*Population Studies Center, University of Michigan, Ann Arbor, MI 48104.

TABLE 1—SIZE AND CHARACTERISTICS OF 50 MUTUALLY EXCLUSIVE RACIAL/ETHNIC GROUPS IN 1980

	Percent			Men 25–54	
	Below Poverty Line	Affluent[a]	Per Capita Income[b]	Mean Yrs. College	Percent Prof./Mgr.[c]
U.S. Total	11	12	$11.2	1.7	28
Hispanic Groups					
Cuban	12	9	10.6	1.7	25
Mexican	22	3	6.8	.7	11
Puerto Rican	34	2	6.1	.6	13
Other Spanish	16	7	8.9	1.5	21
Racial Minorities					
Amer. Natives	25	5	6.9	1.0	19
Asian Indians	8	12	13.2	5.3	60
Black	27	4	7.1	1.0	14
Chinese	13	16	11.9	3.6	42
Filipino	6	9	10.9	2.8	27
Japanese	5	23	14.6	2.9	42
Korean	8	11	10.2	3.4	36
Vietnamese	34	3	5.8	1.4	17
Other Races	21	7	8.8	2.2	26
White Ethnic Groups					
American	10	9	10.0	1.0	18
Amer. Indian	13	6	9.0	.8	15
Armenian	9	19	13.6	2.6	38
Austrian	5	25	15.5	3.3	49
Belgian	7	14	12.3	2.0	36
Canadian	6	7	13.4	2.2	40
Croatian	4	16	13.2	2.1	38
Czech	6	16	13.3	2.2	35
Danish	5	16	13.2	2.3	34
Dutch	8	12	11.7	1.7	29
English	8	15	12.5	2.0	32
Finnish	7	13	11.5	2.1	31
French	8	10	11.6	1.5	27
French-Canadian	6	11	11.7	1.5	25
German	6	14	12.3	1.8	30
Greek	7	16	12.2	2.0	35
Hungarian	6	18	13.5	2.3	36
Irish	8	13	11.8	1.8	29
Italian	6	14	12.1	1.8	31
Lebanese	7	17	13.1	2.8	40
Lithuanian	5	20	14.6	2.6	42
Norwegian	6	15	12.7	2.2	33
Polish	5	11	12.8	2.0	32
Portuguese	7	10	10.9	1.1	19
Rumanian	6	27	15.3	3.3	49
Russian	6	32	17.9	4.0	59
Scandinavian	5	16	12.3	2.7	41
Scots-Irish	5	18	14.1	2.4	37
Scottish	5	20	14.4	2.7	40
Slovak	4	13	13.4	1.8	27
Swedish	5	17	13.7	2.4	37
Swiss	6	18	13.6	2.5	36
Ukrainian	5	17	13.6	2.5	36
Welsh	5	18	13.0	2.6	39
Yugoslavian	7	18	13.7	2.1	33
NEC	6	14	12.4	2.4	36
Not Rpt.	5	11	11.7	2.3	23

Source: U.S. Bureau of the Census, *Census of Population: 1980*, Public Use Microdata Sample.

[a] This reports the percent of individuals in households whose pre-tax cash income equalled at least six times the poverty line.

[b] Shown in thousands: 1987$.

[c] This shows the percent of employed men holding professional or managerial jobs in April 1980.

impoverished, Puerto Ricans and Vietnamese, had rates triple the national average, while among blacks, American Natives, Mexicans, and Other Races, poverty was double the national rate. Only two groups, Vietnamese and Puerto Ricans, were more frequently impoverished than blacks.

Several white groups and one Asian race were overrepresented among the affluent. Nationally, about one person in eight lived in such a household, but one-third of Russians did so. More than one-fifth of Rumanians, Austrians, Lithuanians, and Japanese were in prosperous circumstances.

Per capita incomes are shown in Table 1 in 1987 dollars. Two groups stood out for their meager incomes: Vietnamese (92 percent born outside the United States) had a per capita income just one-half the national average, while Puerto Ricans (about one-half island born) had incomes only $300 above that of the Vietnamese. American Natives, Mexicans, and blacks had small incomes, about 60 percent of the national average. The least prosperous European group (the Portuguese) had incomes exceeding those of blacks by 54 percent.

Russians, whose incomes were $2,500 greater than those of their closest rivals, Austrians and Rumanians, were exceptionally prosperous. Three others (Japanese, Lithuanians, and the Scottish) had incomes $3,000 or more in excess of the overall average.

Information about the educational attainment and occupational achievements of men aged 25–54 is also presented in Table 1. The summary assessment of attainment is per capita years of college education. The national average was 1.7 years, but Asian Indians were exceptional in that they averaged in excess of 5 years of college, or 1.3 years more than the second ranking group, the Russians. Austrians, Chinese, Koreans, and Rumanians had averages exceeding 3 years. At the other extreme were Americans, American Natives, and blacks who averaged just 1 year of college, and 2 groups, Mexicans and Puerto Ricans, whose average was below 1 year.

The final column reports the percent of employed men holding professional or managerial jobs. Overall, 28 percent worked at these positions, but 2 groups stood out for their unusually high proportions: Asian Indians and Russians. In 8 groups, fewer than 20 percent were so employed. Mexicans were least likely to be working as professionals or managers followed by Puerto Ricans, blacks, American Indians, Vietnamese, Americans, American Natives, and one European-origin group, the Portuguese.

III. The Earnings of Men

Does race or ethnicity make a difference? Why are there such substantial discrepancies in economic status? Do some groups invest in education and live in high wage areas, while others quite school and remain in the rural South? Or, are there systematic differences in the rates of financial return groups receive for their human capital investments? A complete disaggregation of differences is beyond the scope of this paper, but groups differ greatly in the characteristics associated with earnings and racial minorities, and Hispanics receive smaller returns for human capital investments than white ethnic groups.

Table 2 describes men aged 25–54 who reported any positive earnings in 1979. To facilitate a comparison, I eliminated men who were foreign born and considered only those 40 groups in which there were 300 or more in the 1980 one-percent sample. In the table, Puerto Rican refers to men born in the United States.

The first column shows mean earnings in 1987 dollars, while the second reports earnings as a percent of the national average of $23,400. There is additional evidence of the prosperity of Eastern European ancestries since Russian, Austrian, and Yugoslavian men had earnings 20 percent or more above average. Other white ethnic groups (with the exception of Americans and American Indians) had earnings 95 to 119 percent of the mean. At the low end were Puerto Ricans, Mexicans, and American Natives, but blacks were at the bottom with earnings 80 percent of the national average. The procedure used here underestimates actual racial-ethnic differences by assuming full employment. In fact, only 83 percent of black men ages 25–54 reported any earnings in 1979, compared to 96 percent of the Russians.

TABLE 2—ANNUAL EARNINGS OF MEN 25–54 AND DIFFERENCES FROM THE NATIONAL AVERAGE FOR 40 MUTUALLY EXCLUSIVE RACIAL/ETHNIC GROUPS IN 1980

			Differences Attributable to	
	1979 Annual Earnings[a]	As Percent of Total	Own Characteristics[b]	Own Rates of Return[c]
U.S. Total	$23.4	100		
Hispanic Groups				
Mexican	19.3	83	−11	−6
Puerto Rican	19.6	84	−8	−3
Other Spanish	21.3	91	−2	−7
Racial Minorities				
Amer. Natives	19.1	82	−12	−12
Black	18.6	80	−12	−8
Chinese	26.8	115	+18	−13
Japanese	26.4	113	+14	−6
Other Races	20.3	87	−1	−16
White Ethnic Groups				
American	21.3	91	−9	0
Amer. Indian	20.8	89	−10	−4
Austrian	29.4	126	+17	+6
Belgian	27.2	116	+7	+8
Canadian	23.7	101	+7	0
Czech	25.5	109	+7	+3
Danish	25.4	109	+9	+2
Dutch	23.5	101	+2	−1
English	24.1	103	+1	+1
Finnish	27.3	117	+5	+5
French	23.2	99	−1	+1
French-Canadian	23.8	102	+2	+2
German	24.4	104	+3	+1
Greek	25.7	110	+10	+1
Hungarian	27.8	119	+10	+8
Irish	24.5	105	+2	+3
Italian	25.5	109	+5	+6
Lithuanian	27.3	117	+11	+3
Norwegian	24.7	106	+6	+3
Polish	25.6	110	+6	+3
Portuguese	25.4	109	+3	−3
Russian	32.4	139	+22	+8
Scandinavian	25.7	110	+8	+2
Scots-Irish	25.4	109	+7	+1
Scottish	25.7	110	+9	+1
Slovak	26.5	114	+9	+3
Swedish	25.6	110	+9	+2
Swiss	24.2	104	+8	−4
Ukrainian	25.5	109	+8	+3
Welsh	25.7	110	+10	+2
Yugoslavian	28.0	120	+9	+15
NEC	24.6	105	+4	+1
Not Rpt.	22.1	95	−7	0

Source: See Table 1.

[a] Shown in thousands: 1987$. Estimates of earnings assume a person worked 2000 hours. Analysis is restricted to all persons reporting positive earnings in 1979.

[b] This is the percent difference between the overall average earnings and the hypothetical earnings of a group assuming their own characteristics and the rates of return of the entire sample.

[c] This is the percent difference between the overall average earnings and the hypothetical earnings of a group assuming their own rates of return but the characteristics of the entire sample.

To investigate causes of these differences, I used an earnings model and a regression standardization procedure. The log of hourly earnings was regressed upon years of precollege education, years of college, marital status (to distinguish married men who earn more), age, and a series of variables that identified the regions and men who lived outside metropolises. Equations were fit for the entire sample and for every group.

For each race-ethnicity, I took the earnings equation for the entire sample and inserted the characteristics of the group to estimate what their earnings would have been if they had the rates of return of the entire sample but their own characteristics. These hypothetical earnings were compared to the national average to assess the net effects of a group's characteristics. They are shown in Table 2 as a percent of the overall mean. Mexicans, for example, had characteristics that cost them an amount equal to 11 percent of the national average.

Next, I used the earnings equations for each group and inserted the characteristics of the total sample. This produced a second hypothetical earnings estimate for every group which was compared to the national average to assess the effects of a group's rates of return. This disaggregation does not unambiguously account for a group's overall difference from the national average.

We expect that groups differ in their earnings because of differences in their geographic locations, education, age, and marital status. Figures in the column Own Characteristics confirm this. The earnings of Russian, Chinese, Austrian, Japanese, and Lithuanian men were raised by 10 percent or more above the average because of their favorable characteristics, while those of Mexicans, American Natives, and blacks were limited because of their characteristics.

Title VII of the 1964 Civil Rights Act outlawed discrimination in the labor market. If this earnings model is appropriate, and if employers comply with the law, we would expect that the numbers showing differences attributable to a group's rates of return would be close to zero. This is the situation for most white ethnic groups. Four white ethnic groups (Belgians, Hungarians, Russians, and Yugoslavians) had earnings that were augmented, but the racial minorities and Hispanics had their earnings limited by lower than average rates of return. The net cost of being an American Native, black, Chinese, or Other Races exceeded 8 percent of the overall mean earnings. It is impossible to demonstrate discrimination using census samples, but these numbers imply that if a specific set of characteristics were possessed by Hispanics or racial minorities, they were less rewarded than if they were possessed by white ethnic groups.

Several additional conclusions may be drawn. First, the high earnings of Eastern European men came about primarily because of their characteristics. Second, Chinese and Japanese reported large earnings but the net effect of membership in these groups was a reduction in earnings. To earn a great deal, Chinese and Japanese must have greater educational attainments than whites (see Ronald Takaki, 1989). Third, the least prosperous of the white groups were those who wrote American or American Indian for their ancestry or left the question blank. Their low earnings were due to their characteristics: their meager educations and concentrations in rural areas. Finally, Mexicans, American Natives, and blacks were at a double disadvantage. Their characteristics limited their earnings, and their rates of return were also below average.

IV. Conclusion

Would the National Academy's Committee have drawn different conclusions if they had considered an array of racial, Hispanic, and ethnic minorities? Is the black-white comparison too simple? The answer to both questions is no. After examining the characteristics of 50 racial-ethnic groups, I found that the Vietnamese and Puerto Ricans were more impoverished than blacks and, in terms of per capita income, Mexicans and American Natives were similar to blacks. There was considerable variation in the status of the 37 white ethnic groups and Eastern European groups had exceptionally high incomes. However, all of the white ethnic groups were more prosperous than blacks. The only white group whose educational attainment or occupational status approximated that of blacks were people who identified themselves as white by race but American Indian by ancestry.

When the analysis focused on earnings of native born men, blacks were the most disadvantaged with earnings below those of Native Americans, Mexicans, and Puerto Ricans. Blacks and men from these three groups had characteristics associated with low earnings. In addition, blacks, other racial minorities (including Asians), and Hispanic groups had rates of return that were lower than those of white ethnic groups.

REFERENCES

Farley, Reynolds, "Race and Ethnicity in the U.S. Census: An Evaluation of the 1980 Ancestry Question," unpublished manuscript, Population Studies Center, University of Michigan, 1989.

Jaynes, Gerald and Williams, Robin, *A Common Destiny: Blacks and American Society*, Washington: National Academy Press, 1989.

Lieberson, Stanley and Waters, Mary C., *From Many Strands*, New York: Russell Sage, 1988.

Snipp, C. Matthew, *American Indians: The First of This Land*, New York: Russell Sage, 1988.

Takaki, Ronald, *Strangers from a Different Shore*, Boston: Little Brown, 1989.

U.S. Bureau of the Census, *U.S. Census of Population: 1980*, PC80-S1-10, Special Report, *Ancestry of the Population by State: 1980*, Washington, 1983.

The Central Role of the South In Accounting for the Economic Progress Of Black Americans

By JAMES J. HECKMAN*

The National Research Council volume *A Common Destiny*: *Blacks and American Society* (Gerald Jaynes and Robin Williams, 1989) synthesizes the fruits of recent research on black social progress since 1940. In a session on this volume held in Atlanta (the cradle of the modern civil rights movement) it is fitting to explore one topical and controversial aspect of black economic progress: the role of government policy and social activism in elevating the status of blacks. The economics profession is currently divided in its assessment of the contribution of federal civil rights and affirmative action policy to the elevation of black economic status. The new NRC report puts together a variety of pieces of evidence that shed valuable new light on this important issue.

One prevailing school of thought emphasizes the importance of long-term secular trends in migration and educational advancement in explaining black economic progress. Another school of thought focuses on the discontinuous improvement in black status that appears in the aggregate time-series of relative black earnings and wages after the passage of Title VII of the Civil Rights Act and initiation of affirmative plans. Two very different views of the labor market experience of black Americans are offered. Does continuous or discontinuous change characterize the recent economic history of black Americans?

The importance of the continuous improvement view to the argument that government played a negligible role is illustrated by the following quotation from James Smith and Finis Welch:

> The racial gap narrowed as rapidly in the 20 years prior to 1960 (and before affirmative action) as during the 20 years afterward. This suggests that the slowly evolving historical forces we have emphasized...education and migration—were the primary determinants of long-term black economic improvement. At best affirmative action has marginally altered black wage gains around this long-term trend.
> [1989, p. 519]

In fact, there is ample evidence of discontinuous change in the improvement of black status during the crucial period 1965–75, even in the studies that claim continuity. The sources of black improvement differ across decades. Migration of blacks out of the South played a major role until 1965, but accounts for much less of the 1965–80 change. The decline in the labor market for the unskilled accounts for the post-1975 stagnation in black relative progress. At issue is what accounts for black progress between 1965 and 1975. Proponents of the continuity hypothesis rely on improvements in the quantity and quality of black schooling during this period. For the trend in black progress to be the same in 1965–80 as it was between 1940 and 1965, the contribution of education and educational quality must have accelerated in this period, because the contribution of migration diminished.

The temporal coincidence of a discontinuous increase in the relative earnings data in the period 1965–75 and the creation of federal policy has led many scholars to assign a central role to government policy in elevating black status. A major difficulty with this argument is that the evidence supporting it is indirect—in large part due to the absence of convincing policy measures. Time-series studies that show an upward jump in the relative status of black earnings after 1964 have convinced few that the measured shift is anything but a temporal coincidence. An apparent enigma that must be explained by those who claim that federal policy was important is that policy agency budgets were small and federal enforcement powers were weak during the period of the greatest black relative wage gains. If enforcement efforts were weak, how could the effect of federal activity have been so strong?

---

*Department of Economics, Yale University, New Haven, CT 06520. This research was supported by NSF grant no. 87-11845. I draw freely from my joint work with John Donahue (forthcoming 1990).

The key to resolving these issues is to disaggregate the data on black economic progress by region and by time period, and to take a broader view of the full range of federal policies developed to eliminate discrimination. Disaggregation of the data reveals the central role of developments in the South for explaining black economic progress. This is hardly surprising since the majority of blacks (the vast majority in 1940) live in that region. Once it is understood that black economic progress in the period 1965–75 is predominately *southern* black economic progress, and that federal policy was directed toward the South in the period, the evidence for a positive policy impact becomes more convincing.

I. Basic Facts of Black Economic Progress: Discontinuity and The Central Role of The South

The aggregate data reveal an upward jump in the time-series of black male earnings and relative wages beginning in the mid-1960s (Richard Freeman, 1973). Much of the jump can be attributed to a jump in the North Central region imposed on a southern upward trend that started before 1965 (Richard Butler and myself, 1978; John Donohue and myself, 1990). Relative improvements in the North Central region die off by 1975. Those in the South do not. The South was the region of the greatest black economic advance in the 1960s. Two-thirds of the growth in black status between 1960 and 1970 occurs there. Blacks made their greatest advance in operative occupations in southern manufacturing (Donohue and myself). There is evidence of substantial desegregation of firms in the South in the period 1966–70 (Orley Ashenfelter and myself, 1976). John Bound and Freeman (1989) use CPS annual March Demographic Files to estimate relative earnings and wage equations for males aged 20–64 over the period 1963–84. Controlling for age, region, urban status, race and education, they document an upward jump in black status after 1965. The period 1966–74 witnessed the greatest black economic gains. After 1975, black status stagnates. Especially noteworthy is the dramatic decline in the percentage wage gap between blacks and whites in the South after 1965 controlling for productivity characteristics. It drops from 38.5 percent in 1965 to 12.6 percent in 1975, accounting for much of the aggregate improvement of black male status. Smith and Welch (1986, Tables A.1 and A.2) corroborate this finding. They find that new entrants to the labor market in the South experienced a dramatic decline in the wage gap between blacks and whites. There is dramatic improvement across all cohorts of workers in the South in the period 1970–80 that Smith-Welch speculate may be a consequence of the fact that "racial discrimination is waning in the South" (1986, p. 49).

The recent growth in black status in the South contrasts with the historical record of relative stagnation. A large body of scholarly literature (beginning with Gunnar Myrdal in 1944 and continuing with Gary Becker, 1957, Donald Dewey, 1952, Robert Margo, 1990, and Gavin Wright, 1988) documents the stability of black-white occupational status in the South over the period 1910–50 in the face of rising relative quantities and qualities of black schooling. Blacks were excluded from new industries and new occupations during this period. Black relative advance of the sort documented by Smith (1984) occurred because of migration to the North. Secular trends of improving black education and advancing industrialization do not explain the stagnant relative status of blacks in the South over this period (Margo). My paper with Brook Payner (1989) documents the stability of patterns of racial exclusion in South Carolina textiles over the period 1910–65. Through two World Wars, the Great Depression, and the Korean War, the share of blacks remained low and stable. A dramatic breakthrough in black employment and wages occurred in the industry only after 1964. The story of blacks in the South is one of stability in relative status up to 1950 in the face of rising relative educational levels and rising industrialization, followed by rapid improvement that was concentrated in the period 1965–75, but began before 1964.

II. Understanding the Change

The central role of black relative improvement in the South in accounting for aggregate black relative improvement in the period 1965–75 helps to resolve some of the controversy concerning the efficacy of government civil rights and affirmative action policy.

One leading explanation of black relative progress advanced by proponents of the continuity hypothesis refers to rising relative levels of black schooling and schooling quality. Smith and Welch have been forceful advocates of this point of view. However, it is important to note that more than 80 percent of their estimated contribution of education to black relative progress is in the improvement in the "return to" (price of) black schooling relative to white schooling during the period 1960–80. The direct contribution of *measured* improvement in years of schooling is generally much smaller for workers in most work-experience classes. Smith and Welch rely on improvement in schooling coefficients as the main source of the contribution of black schooling to black status. Yet improvements in black schooling coefficients may be due to changing discrimination against blacks or changing "quality" of schooling.

There is as yet no microeconomic evidence supporting the claim that increasing black schooling quality raised the "return" (price of) black schooling relative to that of white schooling. Aggregate data on relative schooling expenditures and length of school term in southern segregated facilities show marked improvement in the mid-1940s. Cohorts of black children fully educated in these schools would appear in the labor market in the mid-1960s. Thus there is a real possibility that increasing school quality may account for a substantial fraction of black economic advance in the South in the 1960s.

Before this conclusion is embraced too warmly, however, it is important to note the following qualifications. First, the rapid growth in teacher salaries in southern schools in the 1940s was due to a successful salary equalization program waged by the NAACP (Henry Bullock, 1967). The same teachers were paid more. One form of blatant discrimination was removed by federal court order. Evidence that teacher quality actually improved is weak. The Coleman report (1966) reveals that, by 1965, younger teachers in black schools were relatively inferior compared to their white counterparts than were the older black teachers. Second, the growth in schooling quality for blacks found in the aggregate data may be a consequence of the large outmigration from the rural South where black-white disparities were particularly large. The current micro evidence that growth in relative schooling quality raised black status is weak. Yet, if stronger evidence is found, there will be irony in this for the quality of schooling argument. It was government action in response to NAACP court cases that substantially raised the quality of black schooling.

Direct measures of federal employment policy indicate low enforcement effort during the crucial 1965–70 period. Butler and I and Smith-Welch (1984) note that the greatest black gains occur in a period (1966–72) when the two main civil rights employment and affirmative action enforcement agencies (the Office of Federal Contract Compliance (OFCC) and the Equal Employment Commission (EEOC)) had limited budgets and weak enforcement powers. Many knowledgeable first-hand observers have commented that both agencies were ineffective in their early years. Estimates of OFCC impact summarized in Donohue's and my paper provide no support (pro or con) for the proposition that the agency had a substantial impact on aggregate black relative wages.

Taking a broader view of the law and the measurement of the impacts of the law and recognizing the full range of federal antidiscrimination programs, the case for a positive effect of the law is strengthened. Federal policy was directed toward the South in the crucial period 1965-75. By noting that civil rights policy was directed toward an unwilling South, it is easy to counter the argument that civil rights laws were merely manifestations of preexisting social trends that would have led to black progress under any event. By further noting that black improvement is most rapid in the South, the credibility of a policy effect is substantially strengthened.

The NRC report is particularly helpful in documenting the full range of federal efforts directed toward the South during the period 1965-75. Direct measures of affirmative action and civil rights enforcement activity that have preoccupied previous discussions understate the full thrust of federal activity in the region. The NRC report documents how black voter registration in the South rose from 28 percent in 1962 to 67 percent in 1970. Fifty percent of all charges filed by the Equal Employment Opportunity Commission were filed against firms and establishments in the South during the period 1966-72. Southern schools were the most segregated in 1968 and the least segregated by 1972. These statistics document the full range of federal civil rights activity directed toward the South. In the case of voter registration and school desegregation, they also reveal the success of the policy.

There is evidence that southern employers were eager to employ blacks if given the proper excuse. This produced a strong leverage effect for the new laws. Enforcement budgets could have been small and the law could still have been quite effective. The "enigma" of small budgets and rapid black advance is an artifact of a narrow view of the operation of antidiscrimination laws. My paper with Payner documents that employment of blacks in South Carolina textiles slowed down the growth of labor costs and kept the industry competitive in the period 1965-75 in the face of foreign competition. Integration of geographically isolated textile mills required integration of housing, schooling, and employment. The multipronged federal effort against racial exclusion in all areas of southern life facilitated employment desegregation. It opened new supplies of labor to southern entrepreneurs.

An entire pattern of racial exclusion was challenged. This helps to explain how an apparent straw (the EEOC and the OFCC) could have broken the back of southern employment discrimination. They were only a small part of a large federal effort directed against the South.

Evidence of stability in racial status in the South during the period 1910-50 coupled with dramatic improvement in that region in 1965-75 is consistent with the multiple-equilibrium "community-norms" model of George Akerlof (1980). Wright documents numerous failed attempts at marginal experimentation with the employment of black workers by southern textile firms. Social sanctions apparently played a major role in preserving the southern way of life against marginal experimentation by entrepreneurs. Massive government intervention on a number of interrelated fronts helped move the South from one equilibrium to another, and broke down the racial code. Entrepreneurs were apparently willing to embrace the new laws and make them effective.

However, it is incorrect to attribute all of the black improvement in the South to civil rights laws. The upward trend in black status in the South began before Title VII became law. Social activism in the South combined with improvements in schooling quality and industrial development were also important contributors.

The "quality of schooling" hypothesis does not explain the large cross-cohort improvement in the relative status of blacks in the South in the 1970-80 era (Donohue and myself). Quality of schooling is nothing more than the price paid per unit of schooling. The record from the South indicates that the labor market shifted favorably toward blacks in a fashion that can most economically be accounted for by assigning a major role to federal civil rights policy.

REFERENCES

Akerlof, George, "A Theory of Social Custom, Of Which Unemployment May be One Consequence," *Quarterly Journal of Economics*, June 1980, *94*, 749-75.

Ashenfelter, Orley and Heckman, James, "Measuring the Effect of an Antidiscrimination Program," in O. Ashenfelter and J. Blum, eds., *Evaluating the Labor-Market Effects* *of Social Programs*, Princeton: Industrial Relations Section, 1976.

Becker, Gary, *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.

Bound, John and Freeman, Richard, "Black Progress: Erosion of The Post-1965 Gains in the 1980s?," in S. Shulman and W. Darity, eds., *The Question of Discrimination: Racial Inequality in the U.S. Labor Market*, Middletown: Wesleyan University Press, 1989, ch. 2.

Bullock, Henry, *A History of Negro Education in The South*, Cambridge: Harvard University Press, 1967.

Butler, Richard and Heckman, James, "The Government's Impact on The Labor Market Status of Black Americans: A Critical Review," in L. Hausman, ed., *Equal Rights and Industrial Relations*, Madison: Industrial Relations Research Association, 1978, ch. 9.

Coleman, James, *Equality of Educational Opportunity*, Washington: USGPO, 1966.

Dewey, Donald, "Negro Employment in Southern Industry," *Journal of Political Economy*, August 1952, *60*, 279–93.

Donohue, John and Heckman, James, "Continuous Versus Episodic Change: The Impact of Affirmative Action and Civil Rights Policy On The Economic Status of Blacks," *Journal of Economic Perspectives*, forthcoming, Fall 1990.

Freeman, Richard, "Changes in the Labor Market For Black Americans, 1948–72," *Brookings Papers on Economic Activity*, 1:1973, 246–95.

Heckman, James and Payner, Brook, "Determining The Impact of Federal Antidiscrimination Policy On The Economic Status of Blacks: A Study of South Carolina," *American Economic Review*, March 1989, *79*, 138–77.

Jaynes, Gerald and Williams, Robin, *A Common Destiny: Blacks and American Society*, Washington: National Academy Press, 1989.

Margo, Robert, *Segregated Schools in The American South: A Quantitative History 1880–1950*, Chicago: University of Chicago Press, 1990.

Smith, James, "Race and Human Capital," *American Economic Review*, September 1984, *74*, 685–98.

______ **and Welch, Finis,** "Affirmative Action and Labor Markets," *Journal of Labor Economics*, April 1984, *2*, 269–301.

______ **and** ______, "Closing The Gap," Rand Corporation, Santa Monica, 1986.

______ **and** ______, "Black Economic Progress After Myrdal," *Journal of Economic Literature*, June 1989, *27*, 519–64.

Wright, Gavin, "Segregation and Racial Wage Differentials in the South Before World War II," Working Paper, Stanford University, 1988.

Racial Inequality in the Managerial Age: An Alternative Vision to the NRC Report

By WILLIAM A. DARITY JR.*

Aside from a justifiably dismissive treatment of the usefulness of the "culture of poverty" explanation for the inferior social status of the majority of the black American population (see Gerald Jaynes and Robin Williams, 1989, pp. 540–54), the National Research Council's (NRC) report steers clear of controversy.

The report never mentions the concept of the black underclass, a concept prominent in both popular and academic discussions of the intergenerational persistence of poverty among large segments of the black community. The report ducks altogether the issue of evidence about the construction and execution of genocidal social policies directed toward the black poor. Most fundamental, the report fails to situate the position of the black population within the general matrix of the social class dynamics of American society. Indeed, the understanding of the concepts of social class and class structure displayed in the report is limited to an empiricist social science notion of socioeconomic status, failing to provide a theoretical framework for grasping trends, developments, and future challenges. A quite different analysis is presented here based upon the alternative study prepared by a team of scholars brought together by the William Monroe Trotter Institute at the University of Massachusetts at Boston (see Wornie Reed, 1990).

I

The starting point of the portion of the Trotter Institute study devoted to the economic status of black Americans is the question posed by the sociologist Sidney Willhelm (1970, 1986), "Who needs the Negro?" It is a provocative and disturbing question raising the issue of how nonblacks view the social necessity of the black population in modern America. Willhelm asked this question to highlight dangerous signs that public policy in the United States has taken an exterminative thrust toward black Americans parallel to the genocidal policies directed toward native Americans in the nineteenth century.

In the past, the black American presence in the United States was defined by the need for manual labor. But what happens when the need for such manual labor evaporates? Contemporary developments in the transformation of the class structure of American society reinforces the fears expressed by Willhelm. Dynamic forces affecting the fate of blacks cannot be understood without comprehension of the decisive historical shift that has taken place from capitalist to managerial society. The social dominance of the captains of industry has given way to the social dominance of the intellectuals and the intelligentsia—a professional-managerial elite deeply engaged in the design and administration of social policy (see, for example, my 1986 paper, and Barbara Ehrenreich and John Ehrenreich, 1979). Social status is now more closely associated with the acquisition of credentials and the possession of expert knowledge than with the acquisition of material goods and the possession of wealth. Putative mental capacity supplants property ownership as the key to access to the modern elite.

America today is characterized by a tripartite class structure where the capitalist class and the working class carry over from the previous historical epoch that came to a close in the 1930s with the Great Depression and the New Deal extension of the prerogatives of the state. The present epoch heralds the rise to prominence of a social class that is broadly coterminous with what conventionally is known as the "middle classes"—the managerial class. Under the fading capitalist era, technical change has been motivated by pure profitability considerations, while under the emerging managerial estate, technical change increasingly is charged by pure efficiency considerations. This, in turn, leads not merely to the destruction of occupations associated with the provision of socially necessary labor time, a process characteristic of capitalism, but also the destruction of well-paid, so-called low-skilled and unskilled occupations.

---

*Professor of Economics, University of North Carolina, Chapel Hill, NC 27514.

As even the data from the NRC study (see pp. 272–76) suggest, blacks are disproportionately concentrated in the working class, particularly in that fraction of the working class least utilized by employers. Under capitalist society such a reserve of labor served two functions: 1) it could discipline workers with employment with the threat of replacement by those who formed the pool of the jobless, and 2) the existence of a pool of jobless workers meant that the opening up of new sectors of production could take place without drawing workers from established sectors. But, under managerial society, there is no intrinsic necessity for the reserve of labor. The categories of occupations most accessible to the reserve diminish progressively.

Slavery is legally dead; black workers no longer are needed in droves for gang labor in plantation agriculture in the South. Capitalism's industrial energy is eroding in the United States. It is not clear that there is sufficient employment for white workers. Mechanization and robotization confront a national population experiencing relative deskilling through a school system that has declined in quality over the past 30 years. Coupled with the influx of immigrants from Asia and Latin America, the traditional basis for black labor is reduced further.

Thus, the black working class is placed in an outcast status, simultaneously marginalized and feared by "mainstream" Americans. It is in this context that the concept of the black underclass has meaning: the black underclass is the most degraded and denigrated fraction of the black working class—most frequently without work, without optimism, and highly dependent for income on public assistance, and semilegal and illegal activities. Indeed, it is the black underclass that is widely viewed as in the primary burden on the American system of social services. The NRC report (combining the solicitation of social science experts with substantial government resources) itself appears to be an expression of managerial impulses to construct a consensual solution to "the Negro problem" and "the welfare mess."

II

Aside from certain illusions perpetrated by journals like *Black Enterprise* and *Ebony*, black capitalists are of negligible importance both numerically and in terms of their role in the economy and society at large. "Negro business" as a force in modern America is as mythological as it was when E. Frank Frazier (1957) published his famous study, *Black Bourgeoisie*.

We come then to the paradoxical position of the black members of the rising elite, the members of the black managerial class. Blacks in the professional-managerial class customarily are not positioned at the highest levels of the organizations to which they belong. They rarely play a central role in the creation of policy initiatives. They remain frequently disdained as intellectually inferior by their nonblack peers, and they remain disproportionately concentrated in administrative positions in social welfare agencies (Michael Brown and Steven Erie, 1981).

The widely proclaimed growth in the black middle class in the 1960s and early 1970s associated with claims of "dramatic black progress" were in large measure attributable to the expansion of Great Society programs and the professional employment repercussions at all levels of government (Brown-Erie). These programs played less of a role in generating an increase in the black middle class by uplifting the black poor than by providing direct employment to many blacks as social service providers to other impoverished blacks. Thus, one of the main legacies of the Great Society was to cement the symbiosis between the black poor and the black middle class—the former as the clients of the social service system and the latter as the service providers.

But with the diminished necessity of blacks as laborers, not only is the necessity of the client population increasingly questioned, but also the necessity of the service providers. Thus the black element of the managerial class is the most fragile segment. Some unremarked data presented in the NRC report is indicative of this vulnerability. A table (Jaynes and Williams, p. 273) presenting changes in the occupational and industrial employment distribution for blacks between 1939 and 1984 reveals that, between 1979 and 1984, the proportion of blacks employed in professional occupations fell from 10.7 to 8 percent for men, and fell from 14.8 to 13.9 percent for women. This decline must be attributed to cutbacks in social programs inaugurated under the Reagan Administration that resulted in, as a first effect at the federal level alone, layoffs of black administrators at least at three times the rate of white administrators (Milton Coleman, 1983).

III

The growing perception, particularly among nonblack members of the managerial elite, that blacks are superfluous in American society manifests itself in a wide variety of public policy related measures. These include 1) increased interest in a strategy of preemptive elimination of unborn black babies via population control targeted at the black poor, 2) neglect of health problems concentrated in black neighborhoods, and 3) the tendency to subject black males, in particular, to environments where they can be forgotten or they can die.

While public assistance has undergone a real decline in the 1980s, relatively more resources have been devoted to population control aimed most forcefully at the black underclass. It is not mere coincidence that the first states to fund birth control services on a large scale were in the South in the early 1960s. The southern states have not been known for their "progressive" commitment to a woman's rights to control her fertility. Therefore, southern policymakers appear to have been motivated by concerns over the large black proportions of their state's populations. As Lincoln Pashute reports, "Southern society and its policies [were] motivated, at least in part, by the desire to reduce the fertility of blacks whether for racial reasons or because they believe blacks make welfare demands upon the state in excess of their contribution, or both together" (1978, p. 194).

Recent areal studies indicate that family planning program activity has, indeed, reduced the fertility of low-income black women (John Anderson and Lisa Cope, 1987). Economists Paul Sommers and Laura Thomas have analyzed "the fiscal impact" of the Hyde Amendment by inquiring about "the average cost to the taxpayer of rearing a child from birth to 18 years of age..." (1983, p. 342). They conclude that it would cost considerably less to abort the children that purportedly would prove to be "unwanted births" of low-income women than to provide the children or their mothers with social services. In the same vein, journalist Ken Auletta has been quoted: "the 'incalculable costs' of making abortion harder to obtain. Among those costs: 'More future city criminals will be incubated, unwanted kids, entering the world without nurturing and self-esteem'" (Nat Hentoff, 1989, p. 16). Indeed, before Jesse Jackson began to seek the White House and before he sought to bring the overwhelmingly pro-choice women's movement under the umbrella of the Rainbow Coalition, he "used to say that attempts to make abortion popular in black communities amounted to 'genocide'" (Hentoff, p. 16). Involuntary sterilization of black women, particularly mothers on welfare, and the use of blacks as human guinea pigs for medical experimentation without prior consent has been normal practice in the United States for many years (Thomas Shapiro, 1985).

Neglect of general health conditions in black neighborhoods (higher rates of infection from the AIDS virus, higher rates of infant mortality, higher rates of malnutrition, as well as the destruction wreaked by the illegal drug traffic) all play a direct role in black mortality. The overrepresentation of black males in prisons and as military foot soldiers represents another aspect of the mechanisms for housing or removing those designated as comprising America's surplus population (see my paper with Samuel Myers, 1987).

IV.

The National Research Council's report does not address these matters. A close reading of the report suggests that if only there was a renewal of a period of uninterrupted growth for the American economy, improved public schooling, a renewal of manpower and training programs, and more forceful application of affirmative action measures, blacks would move rapidly toward equality with whites (see especially Jaynes and Williams, p. 319–23). A reformist approach to the evolving social system would prove sufficient for racial convergence from the standpoint of the NRC's study. This is polite, sanitized, and suitable for maintaining the standards of legitimate bureaucratic discourse, but certainly not a realistic assessment of where matters stand.

If blacks are to have a contributing role in the United States, the options are twofold. One option is finally to strike the social compromise that, formally or informally, affords black Americans a strictly representative share of positions in the managerial elite. This would amount to "affirmative action" of a scope unimagined by the NRC. It also would mean that blacks would accept the general inequality that characterizes American society while racial inequality is removed. The second option is to wage a struggle against the social system itself, forming an alliance among all who will be dispossessed with the rise of the managerial class. This means a far more radical course of action and would require calling into question the values that produce success in competitive and hierarchical America.

After all, the most troubling prospect of all is not considered in the NRC report: that a continued black presence is by no means assured. In the transition from capitalism to managerialism, those relegated to the surplus population will find their physical and cultural existence threatened. To the extent that blacks are viewed as *the* social problem (to the extent that a National Research Council will mobilize resources to study "the Negro"), blacks will be the objects of social management, rather than the social system as a whole being viewed as *the* social problem. Not only will the question "Who needs the Negro?" be voiced with greater frequency, but the question "Who is needed and who is not?" will be applied broadly across the entire population. The answers are likely to be as chilling as the question. Indeed, the title of the NRC report is a misnomer. Can there ever be a "common destiny" for all races and all classes in a managerial America?

REFERENCES

Anderson, John E. and Cope, Lisa G., "The Impact of Family Planning Program Activity on Fertility," *Family Planning Perspectives*, July/August 1987, *19*, 152–57.

Brown, Michael K. and Erie, Steven P., "Blacks and the Legacy of the Great Society: The Economic and Political Impact of Federal Social Policy," *Public Policy*, Summer 1981, *29*, 299–330.

Coleman, Milton, "More Reliant on Aid than Whites, Blacks Hit Hard Cuts," *The Washington Post*, December 4, 1983, p. A1.

Darity, William, Jr., "The Managerial Class and Industrial Policy," *Industrial Relations*, Spring 1986, *25*, 212–27.

______ **and Myers, Samuel Jr.**, "Public Policy Trends and the Fate of the Black Family," *Humboldt Journal of Social Relations*, Spring/Summer 1987, *14*, 134–64.

Ehrenriech, Barbara and Ehrenreich, John, "The Professional Managerial Class," in Pat Walker, ed., *Between Labor and Capital*, Boston: South End Press, 1979.

Frazier, E. Franklin, *Black Bourgeoisie*, Glencoe: Free Press 1957.

Hentoff, Nat, "Conundrums and Choices: Jews and Abortion," *Washington Jewish*

Week, October 19, 1989, pp. 14, 16.

Jaynes, Gerald and Williams, Robert Jr., *A Common Destiny: Black and American Society*, Washington: National Academy Press, 1989.

Pashute, Lincoln, "Economics Versus Racial Discrimination in the Provision of Birth Control Services in the United States," in Julian Simon, ed., *Research in Population Economics*, 1978, Vol. 1, 189–96.

Reed, Wornie, *Assessment of the Status of Black Americans*, Vol. 4: *Social, Political and Economic Issues in Black America*, forthcoming 1990.

Shapiro, Thomas J., *Population Control Politics: Women, Sterilization, and Reproductive Choice*, Philadelphia: Temple University Press 1985.

Sommers, Paul M. and Thomas, Laura S., "Restricting Federal-Funds for Abortion: Another Look," *Social Science Quarterly*, March 1983, *64*, 340–46.

Willhelm, Sidney M., "The Economic Demise of Blacks in America: A Prelude to Genocide," *Journal of Black Studies*, December 1986, *17*, 201–54.

______, *Who Needs the Negro?*, Cambridge: Schenkman, 1970.

Are Risk-Benefit Tradeoffs Possible in Siting Hazardous Facilities?

By HOWARD KUNREUTHER AND DOUGLAS EASTERLING*

Ever since Alfred Marshall introduced the notion of consumer's surplus into the literature as a measure of well-being, economists have been interested in determining the amount of compensation required to leave individuals indifferent between the status quo and a new situation where less of a commodity can be consumed (for example, due to an increase in prices). The Hicksian and Kaldor-Hicks criteria of specifying compensation variations may be appropriate for normal market goods where there is a well-specified demand curve. The analysis is less applicable for dealing with environmental resources where elements of risk are involved, or there exist public good features that make it difficult to estimate people's values for different amenities (V. Kerry Smith, 1989).

This paper investigates the question of the appropriate role that providing benefits to a host community (possibly in the form of compensation) can play in improving the chances of siting a facility that is perceived to be potentially hazardous.

In Section I, we develop a simple two-period expected utility model to explain preferences for benefit packages. The empirical analysis examining the predictive power of this model will be related to the following problem of current interest:

The U.S. Congress is attempting to locate a site for storing high-level radioactive waste from commercial nuclear power plants in the United States. An underground repository is to be constructed 2000 feet below the earth's surface for the purpose of permanently storing nuclear waste shipped in casks from power plants around the country. Once filled to capacity, the repository would be sealed to minimize any radiation leakage.

Based on the results of a telephone survey of residents of Nevada regarding their attitudes toward a high-level nuclear waste repository, we suggest ways that the expected utility model might be modified to take into account behavioral factors that are not normally considered part of individual models of choice under uncertainty. These factors may be useful in determining when compensation will be a useful tool in facilitating the siting process.

I. An Expected Utility Model for Siting Facilities

We employ an expected utility model that separates the period in which the siting decision is made from the period during which residents of a community are exposed to risks. Consider an individual who is to receive some level of benefits (B) in return for accepting a nuclear waste repository in his or her backyard. Once the repository is on line, there is a probability p of an accident with negative consequences L. We assume that the person's utility for the option of having

---

[†] *Discussants*: V. Kerry Smith, North Carolina State University-Raleigh; Robert Cameron Mitchell, Clark University; Richard Zeckhauser, Harvard University.

*Riklis Professor, Department of Decision Sciences and Public Policy and Management, Wharton School, University of Pennsylvania, Philadelphia, PA 19104; and Doctoral Student, Department of Public Policy and Management, Wharton School, respectively. We thank Paul Kleindorfer and Marc Knez for helpful comments. The paper is partially based on earlier work undertaken with William Desvousges and Paul Slovic. Research on this paper has been partially supported by NSF grants nos. SES-8809299 and SES86-07111. The telephone surveys are part of a large study contracted to Mountain West and supported by a grant from the State of Nevada.

the repository is given by a two-period additive utility function,

$$(1)\quad U(y,R)=U_1(y_1,R_1)+U_2(y_2,R_2),$$

where $y=(y_1,y_2)$, $R=(R_1,R_2)$, U_t represents the utility of the individual period t, y_t represents income in period t, and R_t represents the consequences of a repository during t. We further assume that each of the U_t satisfies the von Neumann-Morgenstern axioms.

Under the status quo, $R_1=R_2=0$. If the individual accepts a repository with benefits package B provided in period 1 and risk (p,L) existing in period 2, the expected utility is

$$U_1(y_1,B) + [(1-p)U_2(y_2,0)+pU_2(y_2,L)].$$

The level of benefits B^* where an individual is indifferent between maintaining the status quo or accepting the repository is given by

$$(2)\quad U_1(y_1,0)+U_2(y_2,0) = U_1(y_1,B^*)+[(1-p)U_2(y_2,0) + pU_2(y_2,L)].$$

This benefit level B^* represents a willingness to accept (WTA) amount. It is instructive to rearrange equation (2) as follows:

$$(3)\quad U_1(y_1,B^*)-U_1(y_1,0) = p[U_2(y_2,0)-U_2(y_2,L)].$$

The model indicates that an individual's WTA amount should be predicted primarily by the perceived likelihood of a repository accident and the resulting consequences. Assuming U_1 is monotonic in R_1, it also follows that as the level of B increases, more individuals should find the repository an attractive option relative to the present situation. These hypotheses are tested below.

II. Empirical Analysis

This survey was designed to determine the role of the proposed high-level nuclear waste repository at Yucca Mountain, Nevada, which provides an unusual opportunity to examine the importance of different factors in determining individuals' attitudes toward siting a potentially hazardous facility. Two telephone surveys undertaken in early 1987, a national sample of 1,201 U.S. households, and a sample of 1,001 residents of Nevada, examined risk perceptions and attitudes toward the repository. Only data for the Nevada sample are reported here, as the findings from the national sample are extremely similar. The survey was designed to determine the role of various perceptual and attitudinal factors on residents' WTA a high-level repository at Yucca Mountain. Respondents were asked whether or not they would vote in favor of a repository under the condition that an annual federal tax credit would be provided over the next 20 years. The rebate was equal to either $1000, $3000 or $5000, depending upon the experimental condition to which the respondent was assigned.[1]

The empirical analysis that follows investigates how accurately an expected utility model is likely to characterize individuals' attitudes toward voting for a repository. What risk-related factors explain willingness to vote (*VOTE*) for or against a repository? How sensitive are voting preferences to variations in the rebate level?

To answer these questions, we constructed a logit model in which the probability that a person would vote for a repository is a function of a set of independent variables. Duncan Luce and Patrick Suppes (1965) have shown that a logit model is a strict utility model such as (2), under the assumption that the random component of this probabilistic choice model is independent and identically distributed. Analyses of the data indicate that socioeconomic characteristics are not statistically significant in explaining *VOTE* so these factors are not included in the models presented below.[2] The present analyses are based on 470 respondents who responded to the WTA question, and focus solely on how risk-related variables and different rebate levels affect intended voting behavior.[3]

---

[1] This elicitation procedure is a type of referendum model. Robert Mitchell and Richard Caron (1989) have made a strong case for using this type of procedure in contingent valuation surveys.

Impact of Risk-Related Factors. With respect to measuring risk perceptions, we utilized two types of variables: (*i*) proxies for probability and consequences, and (*ii*) more subjective factors related to risk perceptions. With respect to (*i*) the variables are the probability of an accident at the repository (*PR-ACCID*), the probability that repository wastes will leak into the groundwater (*PR-WATER*), the belief that a repository accident would involve certain death (*CONS-DEATH*), and the belief that an accident would kill many people at the time (*KILLMANY*).

Model 1 of Table 1 presents the results. As expected, those individuals who perceive the probability and consequences to be relatively high are more reluctant to vote in favor of the repository at Yucca Mountain. For the $1000 rebate condition, an individual with the most pessimistic values on all four risk predictors has an estimated probability of voting yes equal to .071, while for an individual with the most optimistic values on all four, the estimated probability is .824.

Model 2 of Table 1 examines the incremental effect of four subjective risk measures under category (*ii*). Each of these measures was found to be statistically significant controlling for perceptions of probability and consequences.

The *TRUST* variable refers to the respondent's belief that the federal government will safely manage the repository. The sensitivity of *VOTE* to this factor suggests that for facilities where there is considerable uncertainty regarding the scope of the risk, individuals rely on the credibility of those agents who are seen as controlling the risk.

TABLE 1—LOGISTIC REGRESSION COEFFICIENTS FOR RISK MEASURES IN PREDICTING WILLINGNESS TO VOTE FOR A REPOSITORY WITH REBATES[a]

Predictor	Model 1	Model 2	Model 3
Rebate Level			
$3000	.257	.408	.271
	(0.85)	(1.87)	(0.78)
$5000	.154	.164	.051
	(0.31)	(0.31)	(0.03)
PR-ACCID	−.369	−.298	−.229
	(13.84)^c	(7.47)^c	(4.02)^b
PR-WATER	−.208	.015	.063
	(4.38)^b	(0.02)	(0.28)
CONS-DEATH	−.240	−.136	−.048
	(4.64)^b	(1.25)	(0.14)
KILLMANY	−.212	.004	.036
	(3.36)	(0.00)	(.007)
TRUST	–	.105	.088
		(6.78)^c	(4.54)^b
CONTROL	–	.502	.491
		(13.23)^c	(12.17)^c
DREAD	–	−.269	−.136
		(4.90)^b	(1.10)
RISK-SELF	–	−.162	−.110
		(14.68)^c	(5.86)^b
RISK-FUTURE	–	–	−5.34
			(15.47)^c
Intercept	2.57	.393	.827

[a] The dependent variable *VOTE* is coded as 0 = vote against and 1 = vote for repository. Rebate level is dummy coded with two dichotomous indicators. *RISK-SELF* and *TRUST* are measured on 0–to–10 scales. All others are scored 1–5. *Chi*-square test statistics ($df = 1$) are shown in parentheses.

[b] $p < .05$.

[c] $p < .01$.

Our ability to predict *VOTE* is also enhanced by including two of the subjective risk features identified by Paul Slovic (1987) in the model. *CONTROL* and *DREAD*, respectively, represent the respondent's belief that nearby residents can control the risks of a repository, and that the facility will be dreaded by residents. These two features correspond to the impact of risk exposure per se, rather than the statistical expectation of a loss event. Their significance in explaining *VOTE* suggests that the anticipation of a repository will be anxiety provoking; thus, the benefit package *B* will be offset some-

---

[2] For more details on these findings as well as comparisons between the Nevada and national surveys, see Kunreuther et al. (1988).

[3] A willingness-to-pay question was substituted for the WTA item for 503 respondents. In addition, 28 of those asked the WTA item responded "Don't Know" and were omitted from these analyses.

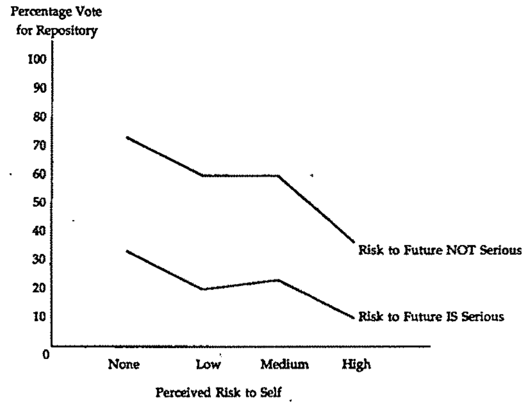


FIGURE 1. EFFECT OF RISK TO SELF AND RISK TO FUTURE GENERATIONS ON WILLINGNESS TO VOTE FOR A REPOSITORY WITH REBATES

what by emotional costs in period 1 of our model.

The fourth indicator of subjective risk perception is an overall evaluation of the seriousness of risk to oneself from the repository. This measure presumably reflects subjective risk features not captured by the other variables in the model. It is highly significant controlling for the other objective and subjective indicators, indicating the difficulty of fully characterizing attitude toward risk.

One of the key issues related to risk perception has to do with the time horizon that respondents find to be relevant in evaluating the acceptability of a nuclear waste repository. In particular, how much weight do individuals assign to accidents or consequences that are likely to occur beyond their own lifetime? This question was addressed by testing the effect of an indicator of risk to future generations (*RISK-FUTURE* in Model 3 of Table 1).

The variable *RISK-FUTURE* was found to have a substantial incremental effect in explaining *VOTE*, indicating that respondents took into account not only their own expected welfare, but also that of their descendants. The independent nature of this effect is evident from Figure 1, where the observed proportion of *YES* votes is plotted as a function of *RISK-SELF* and *RISK-FUTURE*. (Response categories are collapsed for the two risk measures in computing these proportions.)

The size of the *RISK-FUTURE* effect indicates that individuals tend not to discount the future consequences of nuclear waste storage (where the effects are expected to endure for up to 10,000 years), perhaps because of the irreversible features associated with storing high-level radioactive waste in the repository. Confirming evidence on the treatment of future generations comes from a recent study by Ola Svenson and Gunnar Karlsson (1989) who found that a sizeable number of individuals are reluctant to discount to the present the future losses associated with nuclear waste disposal.

Rebate Level. Contrary to our expectations, we found no statistical evidence that preference for the repository differed as a function of the dollar value of the annual rebate. The proportion voting in favor ranged only from 28 to 32 percent across the three levels of $1000, $3000, and $5000 per year.

This finding suggests that the increased acceptability of a repository due to rebates is captured almost entirely in the change from $0 to $1000; the next $4000 in compensation elicits no additional support. Such a result is consistent with a threshold model of choice, whereby individuals refuse to consider compensation if the perceived risk falls in an "inadmissible" range. For those respondents where the risk was perceived to be too high, the rebates offered were viewed *not* as inadequate, but as inappropriate. Given the large proportion of individuals who opposed the repository even at a $5000 annual rebate, it appears that most of the Nevada sample viewed the risks as inherently noncompensable, at least when the benefits are in the form of direct payments to individuals.

III. Discussion

The results suggest that compensation in the form of a rebate is unlikely to have a positive effect on siting a potentially hazardous facility, unless the risk is perceived to be sufficiently low to oneself and to others, including future generations. In the case of nuclear waste facilities, benefits are simply rejected out of hand unless the safety of the facility and the integrity of the siting process are assured (Elizabeth Peele and R. Ellis, 1987; Richard Bryan, 1987). This conclusion does not mean that tradeoffs between risk and benefits are inappropriate when it comes to siting hazardous facilities. Rather it suggests that before one attempts to initiate this process, some threshold level of safety to nearby residents must be assured.

Further, it is clear from the analyses that residents will be confident that the risks are low only if they trust the agencies responsible for constructing and operating the facility; standard risk assessments proffered by experts will not be sufficient. The indispensability of this public confidence was recently recognized by Department of Energy Secretary James Watkins when he abandoned the existing repository planning effort in Nevada. Although these studies had taken two years and cost $500 million, scientific disputes regarding the integrity of the data had fueled effective public opposition to the Yucca Mountain site (Matthew Wald, 1989). The decision delays until at least 2010, and probably longer, the opening of a repository at Yucca Mountain.

REFERENCES

Bryan, Richard H., "The Politics and Promises of Nuclear Waste Disposal: The View from Nevada," *Environment*, October 1987, *29*, 14–17; 32–38.

Kunreuther, Howard, Desvousges, William and Slovic, Paul, "Nevada's Predicament: Public Perceptions of Risk from the Proposed Nuclear Waste Repository," *Environment*, October 1988, *30*, 17–20, 30–33.

Luce, Duncan and Suppes, Patrick, "Preference, Utility, and Subjective Probability," in *Handbook of Mathematical Psychology*, Vol. 3, New York: Wiley & Sons, 1965, 249–410.

Mitchell, Robert and Carson, Richard, *Using Surveys to Value Public Goods*, Washington: Resources for the Future, 1989.

Peele, Elizabeth and Ellis, R., "Hazardous Waste Management Outlook: Are There Ways Out of the 'Not-In-My-Backyard' Impasse?," *Forum for Applied Research and Public Policy*, September 1987.

Slovic, Paul, "Perception of Risk," *Science*, 1987, *236*, 280–285.

Smith, V. Kerry, "Can We Measure the Economic Value of Environmental Amenities?," Presidential Address to Southern Economic Association, 1989.

Svenson, Ola and Karlsson, Gunnar, "Decision-Making, Time Horizons, and Risk in the Very Long-Term Perspective," *Risk Analysis*, September 1989, *9*, 385–98.

Wald, Matthew L., "U.S. Will Start Over on Planning for Nevada Nuclear Waste Dump," *New York Times*, November 29, 1989.

Sources of Inconsistency in Societal Responses to Health Risks

By W. KIP VISCUSI*

Society responds in extreme and often inconsistent ways to health risks. In many instances, the pattern observed is one of overreaction. The Tylenol tampering incidents of the early 1980s drastically reduced the national sales of this product, even though the seven reported deaths were all in the Chicago area. Isolated terrorist incidents periodically choke off the consumer demand for European travel, and the Food and Drug Administration banned the sale of tens of millions of dollars of Chilean fruit based on evidence of low levels of cyanide injected into two grapes. More generally, there is evidence that individuals respond in an alarmist manner to increases in the risks they face, even though these increases may be rather small.

Although one might be tempted to generalize from such events to conclude that there is always universal overreaction to risk, other patterns of behavior reflect errors of the opposite type. Individuals continue to fail to wear seatbelts as often as they should given the health benefits and the costs involved (see Richard Arnould and Henry Grabowski, 1981). Similarly, society has until recently devoted insufficient attention to the long-run environmental problems that we face, including acid rain and the greenhouse effect. Our inaction with respect to these risks can hardly be characterized as a rational response or an overreaction to risk.

There are three possible explanations of such diverse phenomena. First, one could simply dismiss this behavior as being the result of inconsistent and irrational behavior. Second, one could device *ad hoc* explanations of why individuals underreact in some instances and overreact in others. A third possibility is to reconcile this seemingly inconsistent behavior with a consistent theoretical framework. In this paper, I follow the third approach in which I discuss new results that indicate how the character of individual risk perceptions can generate inconsistent patterns of response.

I. The Pattern of Risk Perceptions

The genesis of my approach stems from the relationship between perceived and actual risks. Figure 1 sketches the relationship that has been borne out in studies of risk perception. At probability levels below F_0, individuals tend to overestimate the risk level, whereas for large risks above F_0, there is a tendency toward underestimation. Individuals consequently exaggerate the risks posed by rare events, such as the chance of being hit by lightning, and underassess the truly major risks, such as the chance of death by heart attack or stroke. In addition, the discontinuity of preferences at the zero-risk level indicates that there will be a substantial jump in the perceived risk level once the risk rises from being zero to some nonzero level of risk.

I have incorporated this basic pattern of risk perception within a general theory of decision making under uncertainty, which I have termed "prospective reference theory" (see my 1989 paper). The principal modification in standard choice models is that one replaces the individual's actual probability q with some perceived probability $\pi(q)$. For example, using a *beta* probability distribution, if q is the risk of an accident, then we can write $\pi(q) = (\alpha p + \delta q)/(\alpha + \delta)$, where p denotes the reference risk level, q denotes the risk associated with the particular event, α is the informational content associated with p, and δ is the informational content associated with q.

---

*George G. Allen Professor of Economics, Duke University, Durham, NC 27706.

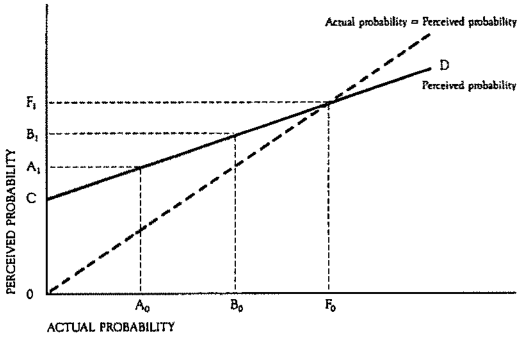


FIGURE 1. RELATION BETWEEN STATED AND PERCEIVED PROBABILITIES.

This formulation is consistent with a Bayesian learning model in which the reference risk p corresponds to the decision maker's prior probability assessment. In the context of accidental death risks, this probability could, for example, be the average risk being addressed in the survey. In the context of laboratory experiments, the probability associated with the lottery outcome could be the probability that would prevail if the respondent did not take the experiment description as being fully informative, but instead placed some weight on a prior in which all outcomes in a lottery were equally likely.

Application of this approach to the anomalies that have been observed in the literature produces quite powerful results. One can reconcile a large and diverse array of types of irrational behavior with the standard expected utility model upon making this transformation. Moreover, what is most striking is that this formulation *predicts* such behavior as opposed to being potentially consistent with these anomalies. For example, the methodology predicts that the Allais Paradox will prevail, that there will be a certainty effect in which individuals value risk reductions that achieve complete certainty more greatly than they should, and many other prominent violations of the standard expected utility theory.

II. Implications for Risk-Taking Behavior

Consider a binary lottery situation in which an individual faces a probability q of injury or death and a probability of $1-q$ of remaining healthy, where the perceived probability is given by $\pi(q)$. In the good health state, utility is given by $U(I)$, and in the ill health state, utility is given by $V(I-L)$, where I is the income level and L is a monetary loss (possibly 0) associated with ill health. For any given level of income, the individual is assumed to rather be healthy than not and to be risk averse. I will also assume that the marginal utility of income is at least as great when healthy as when one is not, which has been borne out empirically in the case of job risks (see my paper with William Evans, 1990).

Let Y be the compensation such as a price cut for a risky product, or a wage premium for a hazardous job that is necessary to maintain the individual's expected utility level at U_0, or

$$(1)\quad U_0 = (1-\pi(q))U(I+Y) + \pi(q)V(I+Y-L).$$

The first tradeoff that will be considered is how the required compensation Y varies with the extent of the loss. Total differentiation of equation (1) yields the result that

$$(2)\quad \frac{dY}{dL} = \frac{\pi(q)V'}{(1-\pi(q))U'+\pi(q)V'} > 0.$$

As the loss increases, the required compensation Y rises. The more important issue is the extent to which dY/dL is altered by the introduction of perceptional biases. For situations in Figure 1 in which q is below F_0, $\pi(q)$ will exceed q, implying that dY/dL will be increased by the biases in risk perception for small risks. Similarly, the required compensation will be decreased by the perceptional biases for high risk levels.

The analogous result for the effect of changes in the risk level q are somewhat more complex, and it is more instructive to consider them within the context of concrete economic actions that will alter the risk level. In particular, suppose that we have opportunities both for insurance and self-protection. In the case of self-protection, one can take a precautionary expenditure c that will influence the risk component q, but the reference risk component p will be unaffected so that we have $\pi(q(c)) = (\alpha p + \delta q(c))/(\alpha + \delta)$. In addition, one can choose to purchase an amount of insurance x for a unit price of s, leading to the optimization problem

$$(3) \quad \underset{c,x}{\operatorname{Max}} Z = (1-\pi)U(I - sx - c) + \pi V(I - L - sx + x - c).$$

The condition for optimal insurance is given by

$$(4) \quad U' = (\pi/q)((1-s)/(1-\pi))V'.$$

Consider the actuarially fair insurance case, where $s = q$. For $\pi(q) > q$, the right-hand side of equation (4) will be larger than in the unbiased case, implying that a lower marginal utility of income in state 2 is needed to establish the optimal insurance amount. A lower marginal utility of income is associated with a higher level of insurance, so that for the risk levels below F_0 in Figure 1, there will be an incentive to overinsure as compared with the unbiased case. For the points above F_0, there will be an incentive to underinsure.

Risk perception biases have a more complex effect on safety precautions. The requirement for optimal self-protection is that

$$(5) \quad -dq/dc = \frac{\alpha+\delta}{\delta}\,\frac{[(1-\pi(q))U' + \pi(q)V']}{(U-V)}.$$

The value $-dq/dc$ indicates the marginal productivity of precautionary behavior in influencing the risk level.

The requirement on the marginal productivity of safety precautions as reflected in the right-hand side of equation (5) is affected in two ways by perceptional biases. The informational content term $(\alpha + \delta)/\delta$, that exceeds 1.0, induces a lower level of precautions by requiring a higher marginal productivity $-dq/dc$. In situations of optimal insurance ($U' = V'$), the net effect will be to diminish the level of precautions selected. When there is not full insurance, the results become more ambiguous because the effect of the risk perception biases on the level of insurance and the expected marginal utility of income hinge on how much marginal utilities are altered when one departs from the optimal insurance amount, and on whether the size of any such effect outweighs the role of the informational weight term. If, however, the dominant effect is that the biases in risk perceptions raise the required marginal productivity of safety expenditures, as in the optimal insurance case, then risk perception biases will always reduce precautionary expenditures.

The nature of the different influences can be summarized using Figure 1. Overall attitudes toward risk and the desirability of insurance will be governed by the relationship between the perceived and actual probabilities. For $q \leq F_0$, risks are overestimated, and there will be a tendency to overinsure and to be overly cautious in discrete responses to risk. For large risks, $q > F_0$, the reaction will be the opposite. From the standpoint of continuous choices affecting safety, however, what is primarily relevant is the slope of CD, not the level of the probability, although this continues to enter the expected marginal utility. Since the perception function $\pi(q)$ flattens out the relationship between perceived and actual probabilities, the marginal efficacy of safety expenditures is reduced. A precaution that reduces a risk from B_0 to A_0 in Figure 1 has a more modest effect of reducing the perceived risk from B_1 to A_1. The risk perception function consequently mutes the perceived impact that safety precautions will have for all levels of risk.

These different competing effects indicate why one might have quite conflicting reactions to the same level of risk. Why, for example, do we respond in often alarmist ways to the various low probability health risks that are called to our attention, yet we fail to take appropriate safety precautions, such as seatbelt use, that are available to us? Although a variety of explanations are possible, the character of the risk perception biases alone is sufficient to explain these seemingly contradictory phenomena. The level of the risk may be overestimated, but the risk perception function may also serve to dampen the perceived efficacy of safety precautions, so that when we have available actions offering incremental reductions in risk, we underrespond.

Although this conclusion is true for marginal changes in riskiness, if there are available strategies that will completely eliminate the risk, then there will be no such dampening in the response. In particular, if we can reduce the risk to zero, we not only obtain a value of the marginal perceived risk reduction probability along *CD*, but we also achieve the additional bonus in terms of the perceived risk reduction of *0C*. Thus, there will be a predilection for policies that achieve the complete certainty of risk reduction. This predilection is borne out in studies of consumer evaluation of product safety, as consumers are willing to pay much more for the final incremental reduction in risk to zero than they are for the earlier risk reductions of equal magnitude, even though economic theory would predict the opposite. Stringent government regulations, such as the Delaney Clause's requirement that no nonzero carcinogenic food additives be permitted, is likewise consistent with this orientation.

The character of the bias is dependent on the nature of the risky decision. Individuals tend to overreact to identified increases in the risk level from its accustomed amount. The study of consumer valuation of product safety by myself with Wesley Magat, and Joel Huber (1987), found that individuals were willing to pay moderate amounts for product risk reductions of 15 injuries per 10,000 bottles of insecticide or toilet bowl cleaner used per year, but when faced with a product risk increase of 1/10,000 most consumers were unwilling to buy the product at all, and those that were demanded a considerable price discount. In this context, the risky choice focused on changes in the risk from the current risk reference point to which consumers had become accustomed. In the case in which consumers are focusing on the risk increase of a product, the jump in the perceived probability indicated by the segment *0C* in Figure 1 is the pertinent perceived risk increase from a marginal shift in the product risk. Even if there were a risk decrease of similar magnitude to the risk increase, there would be no reason to believe that consumers would respond in symmetric fashion, because the reference risk probability p that individuals have with respect to the risks posed by product improvements as opposed to deteriorations in product quality may be quite different.

III. Discounting Deferred Effects

The decision problems that individuals face involving risk are compounded by the task of appropriately discounting these outcomes. Although it has long been speculated that individuals behave myopically, there is no systematic evidence that this is the case. Studies of worker valuations of death risks (such as my paper with Michael Moore, 1989) indicate that the implicit rates of interest with which workers discount the years of life at risk on the job are consistent with rational behavior. Our point estimates of the implied discount rates are in the vicinity of 11 percent, which is somewhat high, but the standard errors on these estimates are sufficient to include other market reference points, such as prevailing mortgage interest rates.

In many respects, examining revealed preferences toward risks to their welfare at different periods of time may represent a best-case scenario. A more important issue from the standpoint of policy is how we will address effects that will not simply influence our own well-being, but also that of our children and future generations. The U.S. Office of Management and Budget (1988) has long specified a 10 percent rate of discount as the main reference point for such calculations—an approach that will drastically reduce the attractiveness of policies such as those that reduce cancer risks, or have long-term implications for our ecological well-being.

Although there have been a variety of battles over the appropriate discount rate, insufficient attention has been paid to the implications of the productivity assumptions that underlie such discount rate estimates. In particular, if the appropriate rate of discount is in fact 10 percent, then the rate of expected productivity growth in the economy also must be quite substantial to justify such a high rate. This growth will boost the income of future generations, which in turn will raise the value that they attach to the risk reduction benefits. Recent estimates by myself and Evans (1990) indicate that the elasticity of the implicit value of job injuries with respect to income is 1.0, and if this relationship generalizes to other health impacts, then it implies that an increase in income will increase the risk reduction benefit values proportionally. Valuing health risks through use of high discount rates should not drastically affect the attractiveness of policies with long-term implications, provided that the benefit values are adjusted appropriately.

Perhaps the main shortcoming is that individuals are likely to place an inefficiently low weight on benefits to future generations. Moreover, our social institutions have thus far proven to be very poor at long range planning, as there is a predilection for responding to more imminent crises. Indeed, if it had not been for the hot summer of 1988, it is unlikely that addressing the greenhouse effect would even be on our national agenda.

As in the case of risk perception biases, the most disturbing aspect of these potential market failures is that the government policies intended to eliminate the shortcomings often appear to be driven by the same set of influences.

REFERENCES

Arnould, Richard and Grabowski, Henry, "Auto Safety Regulation: An Analysis of Market Failure," *Bell Journal of Economics*, Spring 1981, *12*, 27–48.

Viscusi, W. Kip, "Prospective Reference Theory: Toward an Explanation of the Paradoxes," *Journal of Risk and Uncertainty*, September 1989, *2*, 235–64.

______ **and Evans, William,** "Utility Functions that Depend on Health Status: Estimates and Economic Implications," *American Economic Review*, June 1990, 80.

______**, Magat, Wesley A and Huber, Joel,** "An Investigation of the Rationality of Consumer Valuations of Multiple Health Risks," *Rand Journal of Economics*, Winter 1987, *18*, 465–79.

______ **and Moore, Michael J.,** "Rates of Time Preference and Valuations of the Duration of Life," *Journal of Public Economics*, April 1989, *38*, 297–317.

U.S. Office of Management and Budget, *Regulatory Program of the United States Government, April 1, 1988–March 31, 1989*, Washington: USGPO, 1988.

FIXED PREFERENCES AND CHANGING TASTES

By EDI KARNI AND DAVID SCHMEIDLER*

> No man is an island, entire of itself...
> John Donne

The phenomenon colloquially referred to as "fashion" manifests itself to some extent in the consumption of many goods and services, as well as in other aspects of human activity. Thus, we talk about fashionable attire, neighborhoods, furniture, contemporary painters, eighteenth-century composers, as well as fashionable scientific ideas. Although variations in demand caused by fashion affect the risks borne by entrepreneurs and consumers, and the utilization of durable consumer goods (thus exerting important influence on the allocation of resources in the economy), neoclassical demand theory has failed to produce a satisfactory treatment of this phenomenon. This failure may be attributed to the presumption that, by taking individual preferences on the commodity space as given, neoclassical theory is incapable of dealing with phenomena that presumably reflect changing tastes. (Note that while changing tastes due to habit formation may explain certain aspects of variations in individual demands over time, it does not explain the correlated changes in individual demands that constitute a fad.)

In this paper we show that, appropriately extended, neoclassical consumer theory can accommodate the phenomenon of fashion. In other words, within the extended framework, covariations in demand for a given product in the market at large (or among a specific group of customers) that are not accounted for by variations in incomes or relative prices are consistent with fixed individual preferences. The main idea is that the consumption of many commodities is, in part, a social activity. To capture the social aspects of consumption behavior, the standard definition of a commodity, which includes its physical attributes, delivery date, location, and (in the case of contingent commodities) the state of nature, must be extended to include the commodity's social attributes. We claim that the observed patterns of change in the consumption of standard commodities are consistent with constant preferences over the space of extended commodities.

Our concept of extended commodities may be described using the approach of the Austrian School. According to this approach, the utility of bundles of standard commodities is induced by a utility over a space of wants or needs. (See Carl Menger, 1950; Eugene von Bohm-Bawerk, 1959); and for a more recent exposition, Houston McCulloch, 1977.) Standard commodities are consumed as a means of satisfying these wants. In our approach, we distinguish between *material needs* (for example, hunger) and *social needs* (for example, respect, social distinction).

The satisfaction of social needs associated with the consumption of a particular standard commodity depends on the consumption of that commodity and related commodities by other consumers. For example, the satisfaction of a consumer's social needs associated with a pair of brand name sneakers depends on who the wearers of the given brand name are, and who the wearers of other brands of shoes are. The satisfaction of the material need (i.e., the need to protect the feet) depends on the quality of the shoes. The given brand may not differ in this respect from other brands of sneakers. In other words, the social attributes of an extended commodity consist of information concerning the users of this commodity and of information concerning the users of other commodities satisfying the same material needs. Therefore, in terms of the standard commodity space, a consumer's preferences between two standard commodity bundles depend on

---

*The Johns Hopkins University, Baltimore, MD 21218, and Tel Aviv University, Tel Aviv, Israel (Ohio State University), respectively.

the entire allocation in the economy. In the usual terminology, we have a model with consumption externalities. As shown below, the consequences of these externalities for the behavior of demand depends on their specific nature.

To illustrate these ideas, we analyze a dynamic game in which the aforementioned consumption externalities are of the kind that may emerge when there is an exogenous partition of consumers into several (social) classes. The social attributes of a commodity (a vector in this case) are defined by the per capita consumption of the standard commodity in each class; the preferences over standard commodities depend on these social attributes and the class affiliation of the individuals. We show that, in equilibrium, the existence of fixed preferences on the space of extended commodities is consistent with cyclical variations in the demand for standard commodities that resembles the pattern usually associated with fashion. We view the results in this paper less as an end in themselves, but more as a demonstration of the potential of the main idea, and as a first step in the development of a more comprehensive theory that incorporates the social aspects of consumption behavior. (For this reason, we also chose to present our ideas in the context of an overlapping-generations economy. The overlapping-generations structure is not essential for the example below, however, it is a useful dynamic framework that will become handy in further developments of these ideas.)

I. A Three-Color Game

Preliminaries. Consider an economy or a game that evolves through countably many periods, $Z = \{\ldots, -2, -1, 0, 1, 2\ldots,\}$, without a first or a last period. Suppose that there is one good that appears on the market in three different colors, say, red, white, and blue. To highlight the role of the social attributes, we assume that the products under consideration are identical in terms of their ability to satisfy their intended material needs. To avoid unnecessary complications, we assume further that the marginal rate of substitution in production of any two colors is unity, and that this marginal rate of substitution determines the relative prices.

In each period, there is a continuum of players (represented by the unit interval) belonging to distinct social classes, α and β. Every player participates in the game during nine consecutive periods. At the end of each period, one-ninth of the players in each class dies and is replaced at the outset of the following period by the same proportion of players of the same class. One unit of the product is capable of satisfying the consumer's material needs for exactly three consecutive periods, replacing it prematurely is prohibitively costly and replacing it after three periods is mandatory. Thus, each consumer must make a decision on the color of the product every three periods, beginning with the period in which he enters the game. The sequence of choices of all the players defines a *play* of the game. A play of the game is an *equilibrium play* if no player may increase his utility by switching unilaterally to another sequence of choices.

The information available at the end of each period consists of the proportions of the consumers of each class that chose each of the colors at the outset of the same period. For each period $t \in Z$, we denote by α_t^j and β_t^j, $j = 1, 2, 3$, the corresponding proportions, where the three colors are indicated by the numbers 1, 2, and 3.

Social attitudes and payoffs. We assume that consumers' evaluations of a given product's color are determined by the proportion of α consumers and β consumers that use the given color and by their social attitudes. We assume further that social attitudes are, by and large, a reflection of the class affiliation of the individuals. In particular, *ceteris paribus*, (a) the preferences of α consumers for a given color increase monotonically with the fraction of players belonging to their own class that use the same color, and decrease monotonically with the fraction of players belonging to β that use the same color, and that (b) the preferences of β consumers for a particular color increase monotonically with the fraction of players from both classes that use this color. Let the utility function of each consumer be the discounted sum of his one-period utilities over his lifetime,

$$(1)\quad U^i = \sum_{k=0}^{8} \delta^k u^i\big(a^i_{t+k},\ \big(\alpha^j_{t+k}, \beta^j_{t+k} \,|\, j = a^i_{t+k}\big)\big),$$

where $a^i_{t+k} \in \{1,2,3\}$, t denotes the first period in which player i enters the game, and $\delta \in (0,1)$ is a discount factor. Regarding the temporal utility u^i, we assume that for every period τ and $j = 1,2,3$, $u^i(j, \alpha^j_\tau, \beta^j_\tau) = \alpha^j_\tau - 2\beta^j_\tau$ for α consumers and $u^i(j, \alpha^j_\tau, \beta^j_\tau) = \alpha^j_\tau + 2\beta^j_\tau$ for β consumers.

This rather simple model may be applied, with appropriate modifications, in analyzing the demand for housing in an urban area. In particular, the colors in our example may be taken to represent residential neighborhoods, and the class structure the ethnic composition of the population. Regarding the specification of the preferences, John Yinger (1979) reports findings of surveys conducted in the late 1960s indicating that the great majority of whites surveyed object to living in integrated neighborhoods, while the majority of blacks participating in these surveys expressed preferences for living in integrated neighborhoods over living in black neighborhoods. While the expressed preferences may be no more than a reflection of the valuation of the amenities that are associated with neighborhoods populated with specific ethnic groups, the resulting equilibrium may be a constant state of flux as illustrated below.

Existence and characterization of equilibrium plays. We shall now specify a play for the game (a path for the economy) and show that this is an equilibrium play. Let the position of the economy in period t be represented by the triplet of points in the unit square $\{(\alpha^1_t = 0;\ \beta^1_t = 1/3), (\alpha^2_t = 1/3,\ \beta^2_t = 2/3), (\alpha^3_t = 2/3,\ \beta^3_t = 0)\}$. In other words, in period t, one-third of the α consumers consume the product in color 2. Furthermore, one-third of these consumers will die at the end of period t, and, at the outset period $t+1$, the survivors as well as the newly born α consumers will have to choose a color for the next three periods. The remaining two-thirds α consumers consume the product in color 3. They include one-third of the α consumers that had to choose a color at the outset of period t. Similarly, one-third of the β consumers consume the product in color 1 and two-thirds consume the product in color 2. Furthermore, all the β consumers that had to choose a color at the outset of period t (including the newly born) choose color 2, and all those who are currently consuming color 1 will either die at the end of period t, or will have to choose a color at the beginning of period $t+1$. This position is depicted in Figure 1 by the points indicated by 0^j, where for $j = 1,2,3$, (a^j_t, β^j_t) are given by the projections of 0^j.

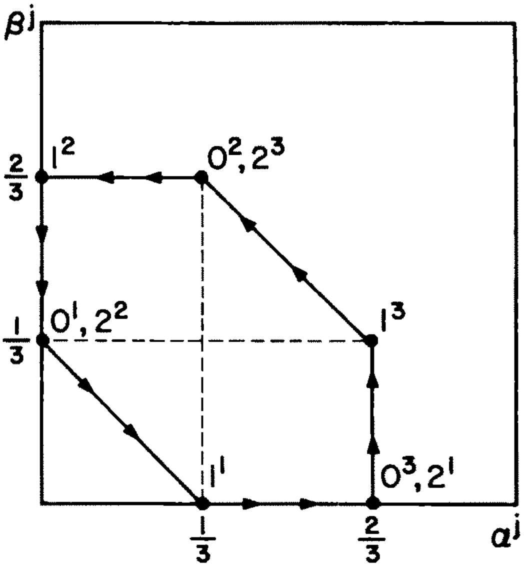


FIGURE 1

We claim that the equilibrium path of the economy (i.e., the path induced by the equilibrium play) includes the positions indicated by $1^j, 2^j$, in periods $t+1, t+2$, respectively. Note that the position depicted by the triplet of points $\{2^1, 2^2, 2^3\}$ in Figure 1 is a rotation of the initial position in which color 1 replaces color 3, color 3 replaces color 2, and color 2 replaces color 1. Clearly, once the position $\{2^1, 2^2, 2^3\}$ has been attained, the next position of the economy is a rotation of the position $\{1^1, 1^2, 1^3\}$ with color 1 replacing color 3, color 3 replacing color 2, and color 2 replacing color 1. The equilibrium path consists of an infinite number of repetitions of these positions in the same order. To show that this path is an equilibrium play, consider the position of the economy at the outset of period t. At this point one-third of the α consumers and one-third of the β consumers must choose the color of their product for the next three periods. Because the utility functions are additively separable, the choice of the most preferred color in each period is independent of previous or subsequent choices. For instance, if i is a new entrant, then he must choose a sequence of three colors, the first to be used in the periods t, $t+1$, $t+2$, the second to be used in the periods $t+3$, $t+4$, $t+5$, and the last to be used during the remaining periods of his lifetime, (i.e., the periods $t+6$, $t+7$, and $t+8$). Given the structure of i's preferences, however, these choices are independent and may be analyzed independently.

The utility corresponding to a choice of color j by player i in period τ is

$$(2)\quad \bar{u}^i(j) \equiv u^i\left(j, \alpha_\tau^j, \beta_\tau^j\right) + \delta u^i\left(j, \alpha_{\tau+1}^j, \beta_{\tau+1}^j\right) + \delta^2 u^i\left(j, \alpha_{\tau+2}^j, \beta_{\tau+2}^j\right).$$

Given the temporal utility functions and taking $\delta = 2/3$, it is easy to verify by direct calculation that the sequence of moves specified by the given play constitutes an equilibrium.

Other than cyclical equilibrium play, the three-color game also has two stationary equilibrium plays: A *uniform equilibrium play* in which, in every period, one-third of the α consumers and one-third of the β consumers choose each of the three colors, and a *segregating equilibrium play* in which each class consumes a different color. Note, however, that if the model is modified to allow for the possibility that individuals have preferences regarding the three colors (say, *ceteris paribus*, red is preferred to white, which in turn is preferred to blue), it is possible to specify games in which the uniform equilibrium does not exist. Furthermore, if the preferences of the β consumers are such that each β consumer would rather use the color that is used by the entire set of α consumers and none of the β consumers than the color that is used by all the β consumers and none of the α consumers, then a segregating equilibrium may not exist. Finally, note that in the present example, the segregating equilibrium Pareto dominates the cyclical equilibrium specified above.

Demand variations. The path of the economy depicted by the cyclical equilibrium play gives rise to cyclical demand for each color reminiscent of variations in demand generally associated with fashion. In addition to the other properties of the model, the specific properties of the cycle depend on the distribution of the population between the two types of consumers. The color life cycle is characterized as follows. Each color displays a life cycle of six periods. In two consecutive periods, say, $t+1$ and $t+2$, a given color is adopted by all the α consumers who must choose a color in these periods, and by these consumers only. In periods $t+3$ and $t+4$, α consumers no longer choose this color, but all the β consumers that must choose in these periods select this color. In periods $t+5$ and $t+6$, no new players of either type choose the given color. At the outset of period $t+7$, the process starts to replicate itself.

The amplitude of the cycle depends on the measure of the two consumer types in the population. If, for instance, the measure of β consumers is twice that of α consumers, then the resulting life cycle of a color will have an amplitude of 2/9. (To see this, note that the measure of β consumers in the population is 2/3, hence when one-third of the β consumers buy the product in a given period, the total demand is 2/9.)

II. Concluding Remarks

The main thesis of our paper is that consumption, taking place in a social environment, is intended to satisfy the social needs as well as the material needs of the consumers. To capture the social aspects of the consumption activity, we propose that the standard specification of commodities in terms of their physical attributes be extended to include their social attributes. Within the extended framework, dynamic variations in the demand for products is an equilibrium phenomenon consistent with fixed preferences and production technology. This analysis suggests that phenomena such as fashion and advertisements aimed at endowing products with a "social image" that could not be adequately analyzed within the framework of neoclassical economic theory, may be explicated using our notion of extended commodities. In addition, a modification that allows the introduction of new product colors is a natural extension of the three-color game. This extension may capture the process of everchanging design of consumer products.

The idea that the demand for standard commodities is induced by their ability to satisfy more basic needs or wants has already been expressed by economists of the Austrian School. More recently, a similar idea appeared in the work of Kelvin Lancaster (1966, 1971), and Gary Becker (1965). Our approach differs from these theories in an essential respect, namely, the recognition of role of standard commodities in satisfying consumers' social needs and the associated externalities. The externalities resulting from the interdependence of preferences in the context of the dynamic general equilibrium framework are responsible for the dynamic variations of demand in equilibrium.

The notion of interdependent preferences is at least as old as modern economic theory (see, for example, Adam Smith, 1937, Book I, ch. XI; Thorstein Veblen, 1899). Harvey Leibenstein (1950) presents a static, partial equilibrium analysis incorporating the Veblen effect and consumption externalities known as the "snob effect" and the "bandwagon effect." More recently, dynamic models incorporating interdependent preferences were studied by Wilhelm Krelle (1973), Wulf Gaertner (1974), and Robert Pollak (1976). The latter two contributions used the framework of linear expenditure systems to formalize the notion of interdependence of preferences. Space limitations preclude a detailed discussion of this literature. (The interested reader will find an insightful discussion in Pollak, 1978.) It should be mentioned, however, that in addition to departing from the linear expenditure system, the approach presented here differs from the previous treatments in an important respect. In previous studies, the utility function of each individual was assumed to depend on the past consumption of other individuals, while the approach presented here is based on simultaneous determination of consumption, thus allowing individuals to anticipate the actions of others. The game theoretic framework adopted here seems adequate to handle this formulation.

Finally, it should be mentioned that George Stigler and Becker (1977) discuss the desirability of modeling fashion with fixed preferences. They regard fashion as a means of achieving social distinction. However, except offering some general observations, they did not develop this idea.

REFERENCES

Becker, Gary S., "A Theory of the Allocation of Time," *Economic Journal*, September 1965, *75*, 493–517.

Gaertner, Wulf, "A Dynamic Model of Interdependent Consumer Behavior," *Zeitschrift fur Nationalockonomie*, 1974, *34*, 327–44.

Krelle, Wilhelm, "Dynamics of the Utility Function," in John R. Hicks and Warren Weber, eds., *Carl Menger and the Austrian School of Economics*, Oxford: Clarendon Press, 1973.

Lancaster, Kelvin J., "A New Approach to Consumer Theory," *Journal of Political Economy*, April 1966, *74*, 132–57.

______, *Consumer Demand: A New Approach*, New York: Columbia University Press, 1971.

Leibenstein, Harvey, "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," *Quarterly Journal of Economics*, May 1950, *64*, 183–207.

Menger, Carl, *Principles of Economics*, Glencoe: Free Press, 1950.

McCulloch, Houston J., "The Austrian Theory of the Marginal Use and of Ordinal Marginal Utility," *Zeitschrift fur Nation-*

alokonomie, 1977, *87*, 249–80.

Pollak, Robert A., "Interdependent Preferences," *American Economic Review*, June 1976, *66*, 309–20.

______, "Endogenous Tastes in Demand and Welfare Analysis," *American Economic Review Proceedings*, May 1978, *68*, 374–79.

Smith, Adam, *The Wealth of Nations*, New York: Modern Library, 1937.

Stigler, George J. and Becker, Gary S., "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, *67*, 76–90.

Veblen, Thorstein, *The Theory of the Leisure Class*, London: Allen & Unwin, 1899.

von Bohm-Bawerk, Eugene, *Capital and Interest*, Libertarian: South Holland, Ill., 1959.

Yinger, John, "Prejudice and Discrimination in the Urban Housing Market," in P. Mieszkowski and M. Straszheim, eds., *Current Issues in Urban Economics*, Baltimore: Johns Hopkins University Press, 1979.

NEW DEVELOPMENTS IN ECONOMIC THEORY[†]

Information as an Economic Commodity

By BETH ALLEN[*]

The words "economic commodities" refer to categories of items that are traded for a price (or, more generally, that can be acquired at some resource cost). Commodities may either be initially present as endowment or be produced from other resources, but economic agents can change their allocations of commodities via trade. A commodity is specified by its complete description, its time and place of delivery, and (for contingent commodities) the event in which it is delivered. In focusing on the commodity features of information (the possibility of costly acquisition), I ignore its very important roles in the theory of incentives (i.e., the principal-agent problem, moral hazard, and adverse selection) and in the study of messages and complexity in resource allocation processes and communication games. Nor will I explicitly analyze those extended Arrow-Debreu economies (as introduced by Roy Radner, 1968) with asymmetric information envisioned as part of the fixed data describing a consumer or firm.

Recently I was pleased to discover that at the annual AEA meeting of exactly 20 years ago, a session was organized around a theme similar to our topic today. At that convention forum entitled "New Ideas in Pure Theory," Radner (1970) discussed equilibrium concepts in the extended Arrow-Debreu model with asymmetric information. He briefly suggested that information could be rendered endogenous in a microeconomic model if one were willing to make the usual assumptions, such as convexity, which would be inappropriate in this context. Several years thereafter, Robert Aumann set forth a formal economic model in which a fixed finite set of information commodities are produced and sold competitively. Needless to say, Aumann's prototype is consistent but he apparently recognized its limitations.

I. Information

For our purposes, economic agents desire information because it helps them to maximize their state-dependent (cardinal) utilities. Formally, let Ω be an abstract set of states of the world and let the collection $\mathbf{F}$ of subsets of Ω represent the measurable events. For simplicity, assume that all agents know the probability μ defined on the σ-field $\mathbf{F}$ of Ω. To avoid the possibility of bankruptcy when consumers do not know their budget constraints accurately, assume that every trader i precisely perceives his strictly positive initial endowment vector $e_i \in \mathbb{R}^l_{++}$. Each agent i understands how his own utility $u_i(x; \omega)$ depends on the true state $\omega \in \Omega$ and the vector x of l ordinary physical (non-information) goods, but the actions of agents do not affect the state of the world. They learn the true event after all trades are concluded, so that decisions can be based only on information available to individuals when the market meets.

Information is technically defined to be a σ-field of measurable events. In other words, an information structure is a generalized partition of Ω with certain additional subsets of Ω added for mathematical convenience. A special case involves information signals correlated with states of the world. Since signals can be viewed as random variables defined on some larger probability space $\Omega \times \Omega'$, they therefore generate σ-fields on $\Omega \times \Omega'$, which can be redefined to equal the set of states of the world in order to transform the familiar signals concept to the framework utilized here.[1]

[†] *Discussants*: David Kreps, Stanford University; Bengt Holmstrom, Yale University; Joseph Ostroy, UCLA.

[*] Professor of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297. Preparation of this paper was supported by NSF grant no. SES88-21442; related research was also supported through grants IST83-14096 and IRI85-20584.

An important aspect of the definition is that information (or, equivalently, an "information structure" for emphasis) involves the entire partition rather than some particular set. Information is the ability to learn a subset in the partition, in contrast to the statement that one particular subset has occurred. In terms of signals, information means the random variable itself (the entire function defined on $\Omega \times \Omega'$) and not a given observation of the random drawing or particular value that the random variable assumes. Economic agents decide whether to acquire information before they can learn the outcome conveyed by the information. (For instance, newspapers must be purchased before reading.)

Since $\mathbf{F}$ always contains more than one σ-field, except in the trivial $\mathbf{F} = \{\Omega, \varnothing\}$ case, different types of information are potentially available. Moreover, consumers generally do not perceive distinct information structures to be equivalent; they may attach more importance to the possibility of learning some events than others. For example, if quality is measured monotonically by a number between 0 and 10, knowing the first digit is much more valuable than the tenth digit after the decimal. Similarly, various traders may evaluate information differently or may wish to obtain different combinations of information structures. For example, an investor needs both price and earnings data on stocks to calculate price-to-earnings ratios, while other traders may focus on alternative indicators such as levels and changes in insider holdings or the total book value of the firm, and chartists want yet different information.

II. Pure Exchange Economies

The preceding argument that various types of information may be available in the economy suggests immediately that we should treat information as a differentiated commodity. Whenever Ω is infinite, many partitions are possible and, in fact, we would generally expect to find uncountably many potential information commodities. Hence, when information is included, the commodity space (which previously was taken to be $\mathbb{R}^l$) becomes infinite dimensional and mathematical problems arise.

To impose some structure, let us consider when different types of information can be close substitutes for each other. Notice first that if they only differ by sets of probability zero, they induce the same economic behavior (when agents' objective functions are maximized conditional on the information) and hence can be regarded as identical for economic purposes. Such information structures are similar if they induce virtually identical expressions for conditional expected utility.

To formalize the notion of similarity, I define the distance between information structures. The metric[2]

$$d(\mathbf{G},\mathbf{H}) = \sup_{G \in \mathbf{G}} \inf_{H \in \mathbf{H}} \mu(G \Delta H) + \sup_{H \in \mathbf{H}} \inf_{G \in \mathbf{G}} \mu(G \Delta H),$$

where $\mathbf{G}$ and $\mathbf{H}$ are complete (i.e., containing all subsets of sets of μ-measure zero) sub-σ-fields of $\mathbf{F}$, defines a topology for information. When information is included in the description of a consumer, the topology respects Werner Hildenbrand's (1974, p. 96) criteria that similar agents should behave similarly in similar situations. Individual demand functions depend continuously on consumers' information, their initial endowment vectors of physical goods, and their concave state-dependent cardinal utilities (see my 1983 paper for details). In addition, the value of information exhibits continuous dependence on its type and we would naturally want the prices associated with information also to depend continuously on the information.

---

[1] This is explained more rigorously in Section 15 of my 1983 article.

[2] Alternatives have been examined in my 1984 paper, and by Kevin Cotter (1986), Maxwell Stinchcombe (forthcoming), and Timothy VanZandt (1988).

The demand for information is a derived demand. Traders desire information for its role in their optimization problems under uncertainty. If this is the only purpose of information, then satiation occurs at one unit of information of any given type.[3] Identical copies of the same information are worthless unless the duplicates can be sold. Moreover, preferences for information are price dependent (and wealth dependent) because pure information is worthless when you cannot afford to make other purchases conditional on the information.

Yet another problem (recognized by Kenneth Arrow, 1962) that arises when one attempts to analyze information as an economic commodity is that information structures are inherently indivisible and useful only in integer amounts. Either you can learn a partition and condition on it or you cannot—half of a particular information structure is meaningless. This introduces fundamental nonconvexities into the model. Due to these indivisibilities, consumers' commodity spaces fail to be convex.

Thus we find that the inclusion of information as an economic commodity violates many standard assumptions in microeconomic theory. Despite these complications, the demand for differentiated information is somewhat well behaved in that it forms a nonempty and upper hemicontinuous correspondence (see my 1986a paper) that (via a finite-dimensional argument) can essentially be convexified by aggregation over uncountably many negligible traders. Furthermore, one can demonstrate that a large perfectly competitive pure exchange economy with endogenous information choices is consistent. There are equilibrium price vectors for ordinary physical goods and continuous equilibrium price functions for information structures (see my 1986b paper).

Cotter analyzes a weaker topology for information and finds that conditional expected utilities and demand functions for physical goods display the same continuity properties. However, the information combination operation (see my 1989 paper with VanZandt) fails to be continuous for Cotter's pointwise convergence metric. Therefore, the endogenous information choice problem must be restricted to the selection of a single information structure by each agent in this framework.

III. Public Goods and Zero Prices for Information

The above discussion has ignored the fact that information can be a public good. Information does not require exclusivity in use —I can read a book and then share it with you without decreasing my information. A more serious public goods problem (which seems to be somewhat more specific to information) concerns the fact that my use of information may reveal it to you. For instance, if you observe my actions and if you suspect that I possess some information not available to you, then my decisions may transmit this information to you. Even when these conditions fail, equilibrium prices can convey information from informed to uninformed traders in a rational expectations model where the initially less-informed agents know the relationship between market prices and states of the world, but need not be aware of the characteristics or choices of other individuals.

A consequence of rational expectations is that traders should be unwilling to invest in the costly acquisition of information if it will be freely revealed by prices. As pointed out by Sanford Grossman and Joseph Stiglitz (1976), this implies that no one would be willing to spend resources to produce information because it can be obtained freely by observing market prices. Yet such information can become extremely valuable when no one else has access to it, so that there is a positive incentive to obtain the information if you are the only person to do so. This paradox means that there may be no equilibrium on the markets for information.

---

[3] In contrast to this "pure" information model, a "hybrid" version in which information can also be enjoyed directly is proposed in my forthcoming paper.

Yet we observe information actually being produced and traded at positive prices. To explain this phenomenon, one can retreat from the pure public goods concept and posit that information commodities retain some private goods features. Thus, traders can benefit from costly information acquisition. (Here, too, a model encompassing the enjoyment of "hybrid" information would help.)

One difficult method of attack requires rational expectations in which prices do not reveal all information. (See my 1986c paper for a survey.) A special case of this approach involves noisy rational expectations, for which Robert Verrecchia (1982) has formulated a partial-equilibrium model of an asset market (with normality and constant absolute risk aversion) for which traders are amenable to the purchase of a single information good at its constant unit cost of production. An alternative strategy explored by VanZandt (1989) focuses on individual-specific information, thereby avoiding the strategic considerations of Section V.

Finally, one can incorporate dynamics into the model and observe that information decays rapidly. Further, rational expectations may not be instantaneous; time is required to learn how to extract new information from market prices, and this timing advantage may suffice to induce some traders to pay to become informed.

IV. Production

To progress from the pure exchange economy to one with production is straightforward for the production of physical goods by firms, but problematic for information production. Not only are infinitely many indivisible differentiated commodities involved, but the technology also should feature large fixed costs, small (perhaps zero) marginal costs over a wide range of output levels, and economies of scope for multiproduct firms.[4] Product differentiation and free entry (or contestability) should lead to monopolistically competitive outcomes.

In the past, information production was explored as part of the economics of research and development. Seminal contributions were made by Arrow and by Jack Hirshleifer (1971), among others. Welfare results and implications for public policy were discussed, although one should note that Hirshleifer formulated a definition of information and pointed out the possibility for informational externalities.[5] The framework discussed here enriches this analysis by permitting various *R&D* projects (that may be complements or substitutes) to be considered, rather than assuming that inventive activity is homogeneous and measurable (usually by input cost) as a single quantity.

The public good characteristics arising from easy or free duplication of information can be recast as a problem of private goods production under extreme economies of scale. Similarly, information fusion (combining several information structures), fission (breaking an information structure apart in order to sell only a piece of it), and the addition of noise can be viewed as elements of information production.

V. Strategic Behavior

I have approached the question of information as an economic commodity from the perfectly competitive paradigm. The hypothesis of price-taking behavior not only serves as a starting point, but also furnishes a useful benchmark for comparison of individual behavior, the existence of equilibrium and its properties, and welfare results in more complicated situations. On the other hand, common sense tells us that information producing firms frequently command considerable monopoly power.[6] Their positions as unique providers can arise from special expertise, patent protection, or copyright laws.

---

[4] My 1990 paper attempts to incorporate such a production sector into a simplified version of the model described earlier.

[5] A paper by Jerry Green (1973) deserves its reputation as a classic for noting that rational expectations create an ambiguity in the sign of the welfare effects of informational externalities because the private benefit from information gathering may or may not exceed its social benefit. Hence either underinvestment or overinvestment in information production may occur.

However, monopolistic behavior need not be confined to firms. If a consumer initially possesses exclusive information, the consequent informational advantage must be considered when deciding to use or sell the information. Vincent Crawford and Joel Sobel (1982) and Morton Kamien et al. (1987) have used noncooperative game theory to explore this issue. The burgeoning recent literature regarding information sharing among firms and the strategic use of information (i.e., in rational expectations models) also relates to the question of whether information can be sold at a positive price. Similarly, signalling models, as originated by Michael Spence (1973), also indirectly address whether costly information investment will be undertaken; the framework here can be used to explain signal choice.

---

[6] Anat Admati and Paul Pfleiderer (1988) have examined a simple model with rational expectations in which a monopolist sells a noisy version of his information.

REFERENCES

Admati, Anat and Pfleiderer, Paul, "Selling and Trading on Information in Financial Markets," *American Economic Review Proceedings*, May 1988, *78*, 96–103.

Allen, Beth, "Neighboring Information and Distributions of Agents' Characteristics Under Uncertainty," *Journal of Mathematical Economics*, September 1983, *12*, 63–101.

______, "Convergence of σ-Fields and Applications to Mathematical Economics," in G. Hammar and D. Pallaschke, eds., *Selected Topics in Operations Research and Mathematical Economics: Proceedings, Karlsruhe, West Germany, 1983*, Berlin: Springer-Verlag, 1984.

______, (1986a) "The Demand for (Differentiated) Information," *Review of Economic Studies*, July 1986, *53*, 311–23.

______, (1986b) "General Equilibrium with Information Sales," *Theory and Decision*, July 1986, *21*, 1–33.

______, (1986c) "General Equilibrium with Rational Expectations," in W. Hildenbrand and A. Mas-Colell, eds., *Contributions to Mathematical Economics in Honor of Gerard Debreu*, Amsterdam: North-Holland, 1986.

______, "The Supply of (Differentiated) Information," in preparation, 1990.

______, "Costly Acquisition of (Differentiated) Information," in R. Parikh, ed., *Theoretical Aspects of Reasoning about Knowledge, 1990*, San Mateo: Morgan Kaufmann, forthcoming.

______ **and VanZandt, Timothy,** "Uniform Continuity of Information Combination: A Corrigendum," mimeo., University of Pennsylvania, 1989.

Arrow, Kenneth J., "Economic Welfare and the Allocation of Resources for Invention," in R. Nelson, ed., *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Universities–NBER Conference Series, No. 13, University Microfilms, 1962.

Aumann, Robert J., "A General Equilibrium Model with Information," mimeo., Department of Mathematics, Hebrew University of Jerusalem, undated.

Cotter, Kevin D., "Similarity of Information and Behavior with a Pointwise Convergence Topology," *Journal of Mathematical Economics*, No. 1, 1986, *15*, 25–38.

Crawford, Vincent P. and Sobel, Joel, "Strategic Information Transmission," *Econometrica*, November 1982, *50*, 1431–51.

Green, Jerry, "Information, Efficiency, and Equilibrium," Harvard Institute of Economic Research Discussion Paper No. 284, 1973.

Grossman, Sanford J. and Stiglitz, Joseph E., "Information and Competitive Price Systems," *American Economic Review Proceedings*, May 1976, *66*, 246–53.

Hildenbrand, Werner, *Core and Equilibria of a Large Economy*, Princeton: Princeton University Press, 1974.

Hirshleifer, Jack, "The Private and Social

Value of Information and the Returns to Inventive Activity," *American Economic Review*, September 1971, *61*, 561–74.

Kamien, Morton I., Tauman, Yair and Zamir, Schmuel, "The Value of Information in a Strategic Conflict," CMSEMS Discussion Paper No. 717, Northwestern University, 1987.

Radner, Roy, "Competitive Equilibrium Under Uncertainty," *Econometrica*, January 1968, *36*, 31–58.

______, "Problems in the Theory of Markets Under Uncertainty," *American Economic Review Proceedings*, May 1970, *60*, 454–60.

Spence, Michael, "Job Market Signaling," *Quarterly Journal of Economics*, August 1973, *87*, 355–74.

Stinchcombe, Maxwell B., "Bayesian Information Topologies," *Journal of Mathematical Economics*, forthcoming.

VanZandt, Timothy, "Information, Measurability and Continuous Behavior," CARESS Working Paper No. 88-18, University of Pennsylvania, 1988.

______, "The Private Acquisition of Information with Individual Risks," mimeo., AT&T Bell Laboratories, 1989.

Verrecchia, Robert E., "Information Acquisition in a Noisy Rational Expectations Economy," *Econometrica*, November 1982, *50*, 1415–30.

Evolution and Cooperation in Noisy Repeated Games

By Drew Fudenberg and Eric Maskin*

The theory of repeated games has been an important tool for understanding how cooperation might arise in a population of self-interested agents. If the prisoner's dilemma, as depicted by the following matrix,

$$\begin{array}{ccc} & C & D \\ C & 2,2 & -1,3 \\ D & 3,-1 & 0,0 \end{array}$$

is played only once, the unique equilibrium is for both players to play D, resulting in the inefficient payoffs (0,0). If, instead, the game is repeated infinitely often and the players are not too impatient, there are "cooperative" equilibria in which both players always play C for fear that failure to do so would cause the opponent to "punish" them with D in the future.

Although repeated play permits cooperation as an equilibrium, it does not preclude less cooperative outcomes. Indeed, it is also an equilibrium for both players to use the strategy "always play D." More generally, the Folk Theorems establish that any feasible individually rational payoffs can arise in an equilibrium if the players are sufficiently patient. (See Robert Aumann and Lloyd Shapley, 1976; Ariel Rubinstein, 1979; and our 1986 article.) In our version of the prisoner's dilemma, a payoff vector is individually rational if each player's payoff is positive.

However, not all these equilibria are equally plausible. Specifically, there is a widespread intuition that often the most likely equilibria are those whose payoffs are efficient, and efficiency is typically *assumed* in economic applications of repeated games. In fact, selecting the efficient equilibria is an approach sometimes taken for games in general (see John Harsanyi and Reinhard Selten, 1988). But the intuition for efficiency seems particularly strong for the case of repeated games.

Herein we provide support for this intuition. For a class of repeated games that includes the prisoner's dilemma, we show that efficiency is implied by evolutionary stability if there is a small probability that the players make "mistakes," so that their realized actions are sometimes different from those they intended.

The evolutionary process we have in mind is one in which pairs of strategies from a population are matched at random to play the repeated game, and a strategy has equal chances of being assigned the role of players 1 and 2. If a strategy performs well on average against the population (i.e., its average payoff is comparatively high), the fraction of the population corresponding to that strategy grows; by contrast, the proportion for a low-payoff strategy declines. In biological settings, such a process results from the relative advantage that a successful strategy has in reproducing itself. In economic or sociological applications, evolution may derive from learning: if players try out strategies on an experimental basis and one works well against the population, more and more players are likely to adopt it (as word gets around), whereas a poor strategy will probably be abandoned.

A strategy is evolutionarily stable (ES) if it can ultimately dominate. That is, following John Maynard Smith (1982), it cannot be "invaded" by any mutant strategy. An ES strategy must be a "best response" to itself, because a mutant playing a better response would have a higher average payoff. Moreover, for strategy s to be ES, there cannot be a mutant s' that does better than s does against a population consisting of a high proportion of s's and a small proportion of s''s.

---

*MIT, Cambridge, MA 02139, and Harvard University, Cambridge, MA 02138, respectively.

Robert Axelrod and William Hamilton (1981) showed that the strategy "always D" is not ES in the repeated prisoner's dilemma with time-average (undiscounted) payoffs. In particular, always D can be invaded by the strategy "tit-for-tat" (that plays C in the first period, and then always plays the same action as its opponent played the previous period). Tit-for-tat can invade because it does as well against always D as always D does against itself (a time-average of 0), and tit-for-tat obtains a payoff of 2 when paired against itself.

Although evolutionary stability rules out always D, the uncooperative outcome can be approximated arbitrarily closely by an ES strategy profile. For example, consider the strategy "Play C in periods $0, n, 2n, \ldots$, and D in all other periods, so long as past play has always conformed to this pattern; if past play has not always conformed, then play D forever afterwards." Call this strategy "mostly D." For large n, mostly D is nearly as uncooperative as always D. Yet it is ES, because if an invader deviates from the pattern, it is punished forever and so obtains a time-average payoff of at most 0. In contrast, mostly D obtains $2/n$ against itself. If the proportion of mutants is sufficiently small, even a mutant that attains an average of 2 when paired against itself cannot overcome this punishment. Thus evolutionary stability by itself has almost no restrictive power in the repeated prisoner's dilemma.

The first step in our argument is the observation that mostly D performs quite badly (payoff 0) against itself if one of the players deviates from the prescribed pattern by "mistake." This suggests that, if mistakes are possible, evolution may tend to weed out strategies that impose drastic penalties for deviations. We then exploit the idea that, if punishments are not too drastic, it is relatively easy for inefficient strategies to be invaded by efficient ones, since the penalty for trying to initiate efficient play against an inefficient opponent is not large.

To formalize the notion of a small probability of mistake, we suppose that each player's expected payoff conditional on there never being a mistake is lexicographically more important than his expected payoff conditional on one mistake, which in turn is more important than his payoff conditional on two mistakes, and so on. In this paper, we consider only the class of symmetric two-player games, which includes the prisoner's dilemma and also the well-known "battle of the sexes." We also assume that players use strategies of only finite complexity: those that can be implemented by a finite computer with a finite memory. We show that if, as in the prisoner's dilemma, there is a unique payoff pair that maximizes the sum of players' payoffs, then any ES (pure) strategy must be efficient. If there are multiple pairs that maximize the sum of the payoffs, then ES does not imply efficiency but still imposes restrictions on the set of equilibrium payoffs.

I. Symmetric Repeated Games with Noise

We consider a symmetric two-player game. Hence, both players choose actions from a common (finite) set A, and we can express a player's payoff as $u(a, b)$, where a is his own action and b that of his opponent. For future reference let $\underline{v} = \min_b \max_a u(a, b)$ be a player's *minmax* payoff and let $V = \text{Convex Hull}\{(v_1, v_2) | \text{there exists } (a, b) \text{ with } u(a, b) = v_1 \text{ and } u(b, a) = v_2\}$ be the set of feasible payoffs.

We are interested in the infinitely repeated game where the above one-shot game is played in each period $t = 0, 1, \ldots$. We suppose that a player's realized action $\hat{a}$ is sometimes different from the action a he intended (so that the realized action is a noisy signal of the one intended), and that payoffs in each period depend only on realized actions. (Alternatively, we could assume that intended and realized actions never differ, but that a player's action might be misperceived by his opponent. In this case, payoffs would depend on both actions and the noise creating the misperception.) Furthermore, a player observes just his opponent's realized action, not the intended one. Thus a player's public history at time t, denoted $h(t)$, is the sequence $[(\hat{a}(0), \hat{b}(0), \ldots, (\hat{a}(t-1), \hat{b}(t-1))]$, where the $\hat{a}$'s and $\hat{b}$'s are his own and his opponent's realized actions, respectively. (Set the history $h(0)$ at date 0 equal to the null set.) History $h(t)$ has length, denoted by $l(h(t))$, equal to t. Let H be the space of all histories. Note that each sequence of play generates two histories, one for each player. For each $h \in H$, let $\pi(h)$ be the corresponding history for the other player, that is, $\pi(h)$ is obtained by permuting $\hat{a}(\tau)$ and $\hat{b}(\tau)$ for each τ. We use this formulation so that, when players use symmetric strategies, we can denote them both by the same map s: $H \to A$. For example, for either player the strategy tit-for-tat in the repeated prisoner's dilemma can be expressed as follows: for all $h \in H$, $s(h) = C$ if $l(h) = 0$; and $s(h) = \hat{b}(l(h)-1)$ if $l(h) > 0$. We will restrict attention to pure strategies that depend only on the public history.

We shall assume that players use strategies of only finite complexity. A formal definition of this concept is introduced by Ehud Kalai and William Stanford (1988). Informally, as we indicated in the introduction, a strategy of finite complexity is a finite program that can be run on a finite computer with bounded memory. Most familiar strategies are of this form, including all the repeated prisoner's dilemma strategies mentioned above.

A pair of finitely complex strategies $\sigma = (s, s')$ gives rise to a sequence of play $[(\hat{a}(0), \hat{b}(0)), (\hat{a}(1), \hat{b}(1)), \dots]$ (the $\hat{a}$'s and $\hat{b}$'s correspond to the players using s and s', respectively) that eventually repeats itself (i.e., forms a repetitive cycle). We suppose that players do not discount, that is, they maximize their time-average payoffs. A player's time-average payoff from σ (contingent on there being no mistake) is

$$v(\sigma) = \lim_{T \to \infty} (1/T) \sum_{t=0}^{T} u(\hat{a}(t), \hat{b}(t)).$$

Let $v(\sigma | h)$ be a player's expected payoff from σ conditional on history h occurring, where $l(h) = \tau$. That is, if $[(\hat{a}(\tau), \hat{b}(\tau)), (\hat{a}(\tau+1), \hat{b}(\tau+1)), \dots]$ is the sequence of action pairs induced by σ after history h,

$$v(\sigma | h) = \lim_{T \to \infty} (1/T) \sum_{\tau}^{T+\tau} u(\hat{a}(t), \hat{b}(t)).$$

To compute the expected payoff conditional on one mistake requires a probability distribution on when mistakes occur. For our results, we can accommodate any distribution with the properties that, conditional on each mistake, each player has a positive probability of making it and each period has a positive probability of being the one in which it occurs. For $m = 1, 2, \dots$, let $v^m(\sigma)$ be a player's expected payoff conditional on there having been m mistakes. By convention, $v^0(\sigma) = v(\sigma)$.

We call $v^m(\sigma)$ a player's "mth-order" payoff from the strategy pair σ. Our lexicographic assumption implies that, for any two pairs σ and $\hat{\sigma}$, if $m < m'$ any difference between $v^m(\sigma)$ and $v^m(\hat{\sigma})$ outweighs a difference between $v^{m'}(\sigma)$ and $v^{m'}(\hat{\sigma})$ in a player's performance. As we mentioned, this formalizes the idea that mistakes are improbable.

The strategies constructed in the papers cited above on the Folk Theorem extend to our "noisy" repeated game, so that for any payoff vector $(v, v) \in V$ with $v \geq \underline{v}$ there is a subgame-perfect equilibrium whose payoffs contingent on no mistakes are (v, v). We will see that considering ES strategies considerably restricts the set of equilibrium payoffs.

II. Evolutionary Stability

As described informally in the introduction, a strategy s is ES if no other strategy s' can invade a population consisting of a large fraction of s's and a small proportion of s''s. We shall suppose that the proportion of s''s, although small, is big relative to the probability of a mistake. Hence, given the lexicographic preferences we have assumed, s' can invade even if it performs worse to the mth-order against s than s does itself, provided that s' fares better to a lower order against itself than s does against s'.

Formally, s is *evolutionary stable* provided that, for all finitely complex strategies s', if there exists m^* such that $v^m(s', s') = v^m(s, s')$ for all $m < m^*$, then there exists $m^{**} \leq m^*$ such that

$$(*) \qquad v^{m^{**}}(s, s) \geq v^{m^{**}}(s', s)$$

and $v^m(s, s) \geq v^m(s', s)$ for all $m < m^{**}$.

Furthermore, if $v^{m^*}(s',s') > v^{m^*}(s,s')$ then $(*)$ is strict.

This definition reduces to the standard one (see Maynard Smith) if the probability of mistakes is literally zero.[1] In that case, only preferences to the 0th order matter, and so s' can invade only if either it performs strictly better than s does against s, or it performs equally well but fares strictly better than s against s'.

We shall say that payoff vector $(v, v') \in V$ is *efficient* if it maximizes the sum of players' payoffs on the set V (i.e., $(v,v') \in \arg\max_{(u,u')}\{u+u' | (u,u') \in V\}$). A strategy s is efficient if $(v(s,s), v(s,s))$ is efficient. Note that the sum of payoffs is relevant because, as we discussed, a strategy has an equal chance of being assigned either player's role.

THEOREM 1: *Let* $\underline{u} = \min\{u | \text{there exists } u' \text{ such that } (u,u') \text{ is efficient}\}$. *If a finitely complex strategy* s *is ES, then* $v(s,s) \geq \underline{u}$. *Thus if there is a unique efficient pair* (u^*, u^*), $v(s,s) = u^*$.

PROOF:

Note first that, for any history h, $(v(s,s|h), v(s,s|\pi(h))) \in V$. Because s is finitely complex, there is a history h^* that minimizes $v(s,s|h)$. We claim that there exists v' such that $(v(s,s|h^*), v')$ is efficient. If not, then in particular $(v(s,s|h^*), v(s,s|\pi(h^*)))$ is inefficient. Choose a sequence of action pairs $\{(a^*(t), b^*(t))\}_{t=0}^{\infty}$ whose time-average payoffs are efficient. Fix actions $a \neq s(h)$ and $b \neq s(\pi(h), (s(\pi(h)), a))$ (where $(\pi(h), (s(\pi(h)), a))$ is the history $\pi(h)$ followed by the action pair $\pi(a, s(\pi(h))) = (s(\pi(h)), a)$. Let s' be a strategy that coincides with s expected after histories h^* and $\pi(h^*)$. After h^*, define s' so that it plays a for two periods. If the opponent plays b in the second period after h^* (i.e., in period $l(h^*)+1$), s' plays the sequence $\{a^*(t)\}$ beginning in the third period after h^* (i.e., in period $l(h^*)+2$, it plays $a^*(0)$, in period $l(h^*)+3$ it plays $a^*(1)$, etc.). If the opponent fails to play b in period $l(h^*)+1$, then s' coincides with s beginning in period $l(h^*)+2$. After $\pi(h^*)$, define s' so that it plays $s(\pi(h^*))$ the first period (i.e., in period $l(\pi(h^*))$). If the opponent plays a in that first period s' plays b in the next period (period $l(\pi(h^*))+1$), and subsequently plays the sequence $\{b^*(t)\}$ (i.e., it plays $b^*(0)$ in period $l(\pi(h^*))+2$, $b^*(1)$ in period $l(\pi(h^*))+3$, etc.). If the opponent fails to play a in period $l(\pi(h^*))$, s' coincides with s thereafter.

We will show that s' can invade. We first claim that

$$(1) \qquad v(s',s|h^*) = v(s,s|h^*).$$

By our choice of h^* and because s' reverts to s after h^* if its opponent does not play according to s', $v(s',s|h^*) \geq v(s,s|h^*)$. Now if this last inequality is strict, let s'' be the strategy that coincides with s' after h^* and otherwise coincides with s. Then $v(s'',s|h^*) > v(s,s|h^*)$, and so if m^* is the number of "mistakes" that must occur if (s,s) gives rise to h^*, $v^{m^*}(s'',s) > v^{m^*}(s,s)$. But by definition of s'', $v^m(s'',s) = v^m(s,s)$ and $v^m(s'',s'') = v^m(s,s'')$ for all $m < m^*$. Thus, s'' can invade, contradicting the evolutionary stability of s. We conclude that (1) must hold.

Because s' plays like s after $\pi(h^*)$ unless its opponent plays according to s',

$$(2) \qquad v(s',s|\pi(h^*)) = v(s,s|\pi(h^*))$$

and

$$v(s,s|h^*) = v(s,s'|h^*).$$

Now, by choice of s', $v(s',s'|h^*) + v(s',s'|\pi(h^*))$ is efficient. Hence, from (1) and because there exists no v' such that $(v(s,s|h^*), v')$ is efficient, $v(s',s'|h^*) + v(s',s'|\pi(h^*)) > v(s',s|h^*) + v(s,s'|\pi(h^*))$. But from (1) and (2) this last inequality is equivalent to

---

[1] Actually it reduces to something slightly weaker than the standard definition. According to the latter, s is ES provided that, for all s', $v(s,s) \geq v(s',s)$ and, if this holds with equality, then $v(s,s') > v(s',s')$. When mistakes have zero probability, we (like Axelrod and Hamilton) deem s to be ES even when the last inequality is weak. However, given the noise in our model, this discrepancy may well be an artifact of our no-discounting assumption.

$$v(s', s'|h^*) + v(s', s'|\pi(h^*)) > v(s, s'|h^*) + v(s, s'|\pi(h^*)). \tag{3}$$

Now (1), (2), and (3) imply that

$$v^{m^*}(s', s') > v^{m^*}(s, s') \tag{4}$$

and $$v^{m^*}(s', s) = v^{m^*}(s, s).$$

Our choice of s' implies, moreover, that

$$v^m(s', s') = v^m(s, s') = v^m(s', s) = v^m(s, s) \text{ for } m < m^*. \tag{5}$$

But from (4) and (5) we infer that s' can invade, contradicting the evolutionary stability of s.

We conclude that there must exist v' such that $(v(s, s|h^*), v')$ is efficient after all. The theorem follows from the definition of h^*.

Because (2,2) is the unique efficient payoff pair in the prisoner's dilemma, Theorem 1 implies that any finitely complex ES strategy must give rise to the cooperative outcome in the repeated game. The qualification "finitely complex" is, however, important. For example, consider the following infinitely complex strategy s^0: "alternate between D and C; if either player deviates from this pattern, switch to one in which C is played only every third period; if either player deviates from *this* pattern, switch to one in which C is played only every fourth period, etc." Strategy s^0 is ES because, regardless of the history, any attempt to deviate from (s^0, s^0) is always punished by having one's payoff reduced by a positive amount. (For a finitely complex strategy, by contrast there must always be a history after which further deviations no longer reduce one's payoff.) But (s^0, s^0) results in the inefficient payoffs (1,1). Indeed, evolutionary stability does not restrict the Folk Theorem payoffs at all if infinitely complex strategies are possible.

Even for finitely complex strategies, Theorem 1 does not necessarily imply efficiency, as we noted earlier. Consider the modified version of the battle of the sexes shown in Table 1. Theorem 1 rules out only payoffs less than 1 in the repeated game (which, however, is still considerably more restrictive than the Folk Theorem, which permits payoffs as low as zero). Indeed, as the following converse of Theorem 1 establishes, any feasible payoff greater than 1 can arise from ES strategies.

TABLE 1—MODIFIED BATTLE OF THE SEXES

	a	*b*	*c*	*d*
a	0,0	4,1	0,0	0,0
b	1,4	0,0	0,0	0,0
c	0,0	0,0	0,0	0,0
d	0,0	0,0	0,0	2,2

THEOREM 2: *Let $\underline{u}$ be defined as in Theorem* 1. *If $(v, v) \in V$ and $v > \max\{\underline{v}, \underline{u}\}$, then there exists a finitely complex ES strategy s such that $v(s, s) = v$, provided there exists a finitely complex strategy with these payoffs.*

Rather than prove this theorem in full generality, we shall simply exhibit ES strategies for the prisoner's dilemma and modified battle of the sexes. In the former, consider the following modification of tit-for-tat: "play C the first period and thereafter play C in a given period if and only if either both players played C or both played D the previous period." Observe that continuation payoffs are always efficient: a mistake triggers a single period of "punishment" and then a return to cooperation. Note that tit-for-tat itself is not ES, because a mistake will trigger the (inefficient) infinite cycle (C, D) (D, C), (C, D) and so on. In the modified battle of the sexes, consider the strategy "Play d the first period and subsequently play d as long as in every past period either both players played d or neither did. If a single player deviates from d, then henceforth that player plays b and his opponent plays a." Even though this strategy is not efficient, it is ES, since any attempt to promote greater efficiency will be punished forever; the punishments are stable because the strategies at that point form an efficient equilibrium all of whose continuation equilibria are efficient.

III. Extensions

In this paper we have assumed no discounting, infinitesimal noise, pure strategies, symmetry across players, and equilibrium configurations where only a single strategy is played. The intuitive reasons behind our results, however, seem quite general, and so in a forthcoming paper we expect to relax all these assumptions. Our work builds on an extensive literature, but, for lack of space, we must also postpone discussion of previous work to the sequel.

REFERENCES

Aumann, Robert and Shapley, Lloyd, "Long-Term Competition: A Game-Theoretical Analysis," mimeo., 1976.

Axelrod, Robert and Hamilton, William, "The Evolution of Cooperation," *Science*, November 1981, *211*, 1390–96.

Fudenberg, Drew and Maskin, Eric, "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, May 1986, *54*, 533–56.

Harsanyi, John and Selten, Reinhard, *A General Theory of Equilibrium Selection in Games*, Cambridge: MIT Press, 1988.

Kalai, Ehud and Stanford, William, "Finite Rationality and Interpersonal Complexity in Repeated Games," *Econometrica*, March, 1988, *56*, 391–410.

Maynard Smith, John, *Evolution and the Theory of Games*, Cambridge: Cambridge University Press, 1982.

Rubinstein, Ariel, "Equilibrium in Supergames with the Overtaking Criterion," *Journal of Economic Theory*, January, 1979, *21*, 1–9.

The Economic Implications of an Incomplete Asset Market

By H. M. POLEMARCHAKIS*

When the asset market is incomplete, the role of prices extends beyond conveying the aggregate scarcity of commodities. In conjunction with the asset structure, they determine the attainable reallocations of revenue. This affects nontrivially the existence, optimality, and determinacy of competitive equilibrium allocations, as well as the revelation of information by prices. Further, it accounts for diverse phenomena, among them the preservation of memory in macroeconomics aggregates.

A simple exchange economy extends over two periods. Uncertainty, indexed by finitely many states of nature, $s=1,\dots,S$, is resolved in the second period. Commodities, $l=1,\dots,L$, are traded in spot markets in the second period after the uncertainty has been resolved and assets have paid off. A commodity bundle is $x=(\dots,x(s),\dots)=(\dots,x_l(s),\dots)$. Commodity prices are $p=(\dots,p(s),\dots)=(\dots,p_l(s),\dots)$.

Assets, $a=1,\dots,A$, are traded in the first period and pay off in the second. A portfolio is $y=(\dots,y_a,\dots)$. Assets are real: the payoff of an asset is a commodity bundle $r_a=(\dots,r_a(s),\dots)$. At commodity prices p, the payoff of an asset, more precisely the payoff of an asset in terms of revenue, is

$$r_a(p)=p'\square r_a=(\dots,p'(s)r_a(s),\dots).$$

For example, the payoff of a futures contract in commodity l is $r_a(p)=(\dots,p_l(s),\dots)$.

The asset structure is the matrix $R=(\dots,r_a,\dots)$. Asset payoffs are $R(p)=(\dots,r_a(p),\dots)$.

A reallocation of revenue is $\tau=(\dots,\tau(s),\dots)$. The attainable reallocations of revenue coincide with the column span of the matrix of asset payoffs, $[R(p)]=\{\tau\colon \tau=R(p)y$, for some portfolio, $y\}$.

An individual expresses excess demand for commodities, $z=(\dots,z(s),\dots)$, so as to maximize utility. The reallocation of revenue necessary to finance the excess demand is $p'\square z=(\dots,p'(s)z(s),\dots)$. An individual faces the attainability constraint $p'\square z\in[R(p)]$, in addition to the budget constraint $p'z=0$.

The attainable reallocations of revenue depend nontrivially on commodity prices. For example, if the asset structure consists of forward contracts in two commodities, and if the relative prices of these two commodities coincide in two states of nature:

$$p_l(s)/p_{l'}(s)=p_l(s)/p_{l'}(s'),$$

independent reallocations of revenue into these two states are not attainable.

The asset market is complete if and only if all reallocations of revenue are attainable; otherwise it is incomplete. Evidently, for the asset market to be complete, it is necessary but not sufficient that $A=S$. At competitive equilibrium prices, the commodity markets clear; the asset market clears as a residual. Associated with competitive equilibrium prices there is an allocation of commodities; also an allocation of assets.

Competitive equilibrium allocations typically fail to be constrained Pareto optimal. Since individuals optimize under the attainability constraint, it is not surprising that competitive equilibrium allocations typically fail to be optimal. What is surprising is that competitive allocations typically fail to satisfy an appropriate criterion of constrained optimality.

At a given allocation of assets, the attainability constraint takes the form $p'\square z=R(p)\bar{y}$, where $\bar{y}$ is the portfolio assigned to the individual. At competitive equilibrium prices for a given allocation of assets, commodity markets clear. Associated with competitive equilibrium prices for a given allocation of assets, there is an allocation of commodities. A competitive equilibrium allocation of commodities is constrained pareto suboptimal, if and only if there exists a competitive equilibrium allocation for a given allocation of assets that Pareto dominates it.

---

*Core, Université Catholique de Louvain, 1348 Belgium, and School of Business, Columbia University.

When the asset market is incomplete, competitive equilibrium allocations of commodities are typically constrained suboptimal. The reason is as follows. A reallocation of assets and the resulting reallocation of revenue across individuals and across states of nature have two effects on the welfare of an individual: a direct, revenue effect; and an indirect relative prices effect due to the change in relative prices required to attain market clearing following the reallocation of revenue at each state of nature. Since at a competitive equilibrium individuals optimize in the asset market, the revenue effect following a reallocation of assets cannot be Pareto improving. Nevertheless, when the asset market is incomplete, the marginal rates of substitution of revenue across states of nature for different individuals typically do not coincide. The relative price effect may then yield a Pareto improvement. (See John Geanakoplos and myself, 1986.)

Constrained suboptimality imposes on a policymaker who intervenes in the asset market the constraint to employ only the instruments available to the market, the existing assets. The claim of constrained suboptimality is strengthened by the result that the observable behavior of individuals (their demand functions in the asset market and the spot commodity markets), suffices in order to recover the utility functions of individuals even when the asset market is incomplete and hence to determine improving interventions. (See Geanakoplos and myself, 1990.)

Competitive equilibria may not exist. This is due to the discontinuous dependence of the attainable reallocations of revenue on commodity prices. For example, if the asset structure consists of forward contracts in two commodities, discontinuities occur at commodity prices at which the relative price of these two commodities coincide at all states of nature:

$$p_l(s)/p_{l'}(s) = p_l(s')/p_{l'}(s'), \text{ for all } s, s'.$$

At such prices, the two-dimensional set of attainable reallocations of revenue changes discontinuously to a one-dimensional set. The resulting discontinuity in demand may cause competitive equilibria not to exist. Nevertheless, for a family of economies indexed by the asset structure and the endowments of individuals, competitive equilibria typically do exist. (See Darrell Duffie and Wayne Shafer, 1985.)

Alternatively, assets are nominal: their payoffs are denominated in abstract units of account, inside fiat money in zero net supply.

The prices of units of account are $\delta = (\dots,\delta(s),\dots)$; equivalently, this is the purchasing power of units of account. The payoff of an asset in units of accounts is $r_a = (\dots, r_a(s),\dots)$. The payoff of an asset is $r_a(\delta) = \delta' \Box r_a$. The matrix of asset payoffs is $R(\delta) = (\dots, r_a(\delta),\dots)$. The attainable reallocations of revenue are $[R(\delta)] = \{\tau\colon \ \tau = R(\delta)y$, for some portfolio, $y\}$.

The attainability constraint takes the form $p' \Box z \in [R(\delta)]$. At competitive equilibrium prices for commodities and units of account, the markets for commodities and units of account clear; the asset market clears as a residual. Associated with competitive equilibrium prices, there is an allocation of commodities; also of units of account and of assets.

In order to eliminate nominal differences across equilibria, let some commodity be numeraire, $p_{l*}(s) = 1$, for all s. Since units of account are in zero net supply and provide no liquidity or other services, the market for units of account clears at all prices. Evidently, the price of units of account or their purchasing power can be normalized to equal one at each state of nature; then, the budget constraint at each state of nature is not homogeneous.

Indeterminacy occurs, and neutrality (the dichotomy between real and nominal variables) fails when competitive equilibrium commodity allocations associated with distinct prices for units of account are distinct. When the asset market is complete, indeterminacy does not occur. This is evident since all reallocations of revenue are attainable and the attainability constraint vanishes.

Suppose, for simplicity, that asset $a=1$ pays off one unit of account in state $s=1$ and zero in every other state, while assets $a=2,\dots,A$ pay off zero in state $s=1$; equivalently, consumption in state $s=1$ can be interpreted as consumption during the first, contracting period.

When the asset market is incomplete, variations in the prices of units of account in states $s=2,\dots,S$ that are not colinear typically alter the attainable reallocations of revenue across these states. Since the individual's holdings of assets $a=2,\dots,A$ are not constrained by the budget constraint, the associated competitive equilibrium allocations of commodities are distinct as well; this is the case provided individuals are sufficiently numerous and diverse. (See David Cass, 1985.) The degree of indeterminacy, the dimension of a set of distinct competitive equilibrium allocations of commodities in $S-2$. (See Geanakoplos and Andrew Mas-Colell, 1989.)

Indeterminacy occurs even when the unit of account is outside money in positive net supply that provides liquidity services. It suffices that individual precautionary demand for money balances be determined before the resolution of uncertainty in the second period. The indeterminacy of competitive equilibrium commodity allocations translates into the indeterminacy of asset prices in the first period. When individuals are differentially informed, rational expectations equilibria may be completely noninformative. (See my 1990 paper with Paolo Siconolfi.)

Consider an exchange economy of overlapping generations. A single perishable consumption good is available each period. An aggregate individual is born each period and lives for three periods. The economy is stationary: individuals are identical other than for the calendar time of their birth. The endowment of the individuals in the three periods of his life is $(\theta,1,0)$, and his intertemporal, von Neumann-Morgenstern utility function over consumption in the three periods of his life, (x_1,x_2,x_3) is $\ln x_2+\delta\ln x_3$, $1>\delta>0$.

The endowment of the aggregate individual in the first period of his life is a positive random variable that is i.i.d. in time according to the frequency function $\pi(\theta)$. It follows from the time invariance and intertemporal separability of individual preferences and of the aggregate shock that an optimal and time-invariant allocation is i.i.d. in time; aggregate output is

$$\begin{aligned} y_t^* = y^*(\theta_t) &= \theta_t+1-x_2(\theta_t) \\ &= (\delta/(1+\delta))(1+\theta_t). \end{aligned}$$

This is a competitive equilibrium allocation if the asset market is complete: before economic activity commences, elementary securities are traded for the transfer of revenue across time periods and realizations of uncertainty; all individuals, irrespective of the calendar time and realization of uncertainty of their birth participate in this initial exchange.

Alternatively, with outside fiat money in fixed supply, M, as the store of value, at a time-invariant allocation aggregate output is

$$\begin{aligned} \hat{y}_t &= \hat{y}(\theta_t,\hat{m}_t) \\ &= \frac{(\theta_t+\delta+\theta_t\delta)(M-\hat{m}_t)}{M+\delta(M-\hat{m}_t)}, \end{aligned}$$

where $\hat{m}_t$ is the money held by individuals in the second period of their lives and $M-\hat{m}_t$ by those in the third period. Evidently, since $\hat{m}_t=\hat{p}_{t-1}\theta_{t-1}$, aggregate output is serially correlated; so is the price level,

$$\begin{aligned} \hat{p}_t &= \hat{p}(\theta_t,\hat{m}_t) \\ &= \frac{M+\delta(M-\hat{m}_t)}{\theta_t+\delta+\delta\theta_t}. \end{aligned}$$

Also, a Philips trade off exists between the deviation of output from its optimal level and the rate of inflation. (See Jayasri Dutta and myself, 1990.)

Here, the incompleteness of the asset market is primitive. It is an open question to derive the incompleteness of the asset market from primitive differences of information among individuals.

REFERENCES

Cass, David, "On the 'Number' of Equilibrium Allocations with Incomplete Financial Markets," mimeo., 1985.

Duffie, Darrell and Shafer, Wayne, "Equilibrium with Incomplete Markets, I: A Basic Model of Generic Existence," *Journal of Mathematical Economics*, No. 1, 1985, *14*, 285–99.

Dutta, Jayasri and Polemarchakis, Heraklis, "Asset Markets and Equilibrium Processes," *Review of Economic Studies*, forthcoming, 1990.

Geanakoplos, John and Mas-Colell, Andreu, "Real Indeterminacy with Financial Assets," *Journal of Economic Theory*, February 1989, *47*, 22–38.

______ **and Polemarchakis, Heraklis,** "Existence, Regularity and Constrained Suboptimality of Competitive Allocations when the Asset Market is Incomplete," in Walter Heller et al., eds., *Uncertainty Information and Communication: Essays in Honor of Kenneth Arrow*, Vol. 3, New York: Cambridge University Press, 1986.

______ **and** ______, "Observability and Optimality," *Journal of Mathematical Economics*, forthcoming, 1990.

Polemarchakis, Heraklis and Siconolfi, Paolo, "Noninformative Rational Expectations Equilibria," mimeo., 1990.

Child Care and Women's Return To Work After Childbirth

By JACOB ALEX KLERMAN AND ARLEEN LEIBOWITZ*

Over the last two decades, the need for nonmaternal child care has risen dramatically as increasing numbers of mothers with preschool children have entered the labor force. In 1960, most women left the work force during pregnancy and remained out of the labor force at least until their youngest child entered school. Today, more than one-half of all mothers return to work before their youngest child's second birthday (U.S. Bureau of the Census, 1987).

Our work suggests that growth in wages and declines in family income apart from the wife's earnings both contributed to the rise in new mothers' labor supply over time (see our 1988 paper with Linda Waite). This paper extends that work to consider the effects of child care costs.

As the percentage of women working when their children are young has increased, the importance of child care costs in the family's decision making has also grown. Child care costs loom especially large for families with children under 2-years old for two reasons: first, child care for infants is more labor intensive and costs more than child care for older preschoolers (Martin O'Connell and David Bloom, 1987); second, for children aged 5 and above, schools provide "free" child care for at least part of the day.

In recent years, several federal and state programs have subsidized the cost of child care. Income tax deductions and credits for child expenses form the major source of federal subsidy to child care (Philip Robins, 1988). This paper examines the role of these subsidies in the observed rapid increase in work among mothers of very young children.

Previous work, aggregating data for all mothers regardless of the age of their children, has found several indicators of child care cost that are correlated with women's work: Another adult (other than the husband) living in the household, living in areas with lower average child care costs and more preferential tax treatment of child care expenses (James Heckman, 1974; Rachel Connelly, 1989; David Blau and Robins, 1989). In this paper, we focus on the labor supply of women immediately following their first birth in order to understand to what extent child care subsidies have promoted the rapid growth in labor supply that has been observed among mothers of very young children.

We briefly discuss a model of women's return to work and our implementation of that model using data from the *National Longitudinal Survey of Youth* (*NLS-Y*). Empirical estimates of the effect of child care costs on return to work after childbirth follow.

I. The Model

After childbirth, the presence of an infant keeps the value of the mother's time in the home at a high level. But, as children progress from newborn to toddler to preschooler, they

---

[†] *Discussants*: Rachel Connelly, Bowdoin College; Kathryn Anderson, Vanderbilt University; Robert Topel, University of Chicago.

*Economics and Statistics Department, The Rand Corporation, 1700 Main Street, Santa Monica, CA 90406-2138. This research was funded by grants HD-20253 and P-50-HD-12639 from the National Institute for Child Health and Human Development and by grant J-9-J-7-0091 from the Department of Labor. We are indebted to David Blau for supplying data on child care credits, to Sally Carson for programming assistance, to Linda Waite for helpful comments, and to Kathryn Vitali for her help with the manuscript.

require less-intensive inputs of their mother's time. As the child grows, the marginal product of the mother's home time falls, and work becomes increasingly attractive. Because women vary in the market wage available to them, in their home productivity, and in the price they face for child care, the timing of their return to work will also vary.

Historically, relatives have been a major source of child care subsidy, since they often provide free or lower-cost care. Care by relatives is also highly valued by parents, particularly for young children (Karen Mason and Karen Kohlthau, 1989). The presence of family members in the household who can substitute for the mother in child care reduces the marginal product of her time at home and should promote earlier return to work.

Income tax credits for child care costs also increase the net wage for any given market wage. We hypothesize that both relatives and lower-cost child care (through lower wages for child care workers or subsidies) will reduce the interval between the birth and the date at which the home productivity is exceeded by the market wage, and therefore be related to earlier return to work. However, child care tax credits subsidize market-provided child care, but not that provided by unpaid relatives, thus altering the relative prices of market and nonmarket-provided care. We hypothesize that tax subsidies are more important determinants of return to work with market-provided child care than return to work with care by relatives.

II. Data and Statistical Methods

This analysis uses data collected by the *NLS-Y*, a panel survey of nearly 6,000 young women that includes detailed information on labor market behavior. The panel began in 1979 with women aged 14 to 21. We have used the subsequent interview data to construct a history of labor supply following each first birth that occurred between 1979 and 1987. Because we have data only through 1987, when the women were aged 22–29, this sample of first births overrepresents women who gave birth at relatively young ages.

Here we estimate multinomial logit models of the determinants of returning to work by 3 and 24 months following childbirth and the mode of child care utilized (market child care, i.e., day care center or other care by a nonrelative; or nonmarket care, i.e., relative, paid, or unpaid). There are several reasons for estimating multinomial logit models: First, computational convenience dictated their use.[1] Second, estimating multinomial logits at two points in time allows the determinants of the decision to work and the type of child care utilized to vary, as the time since childbirth increases. This is important since the type of child care most appropriate changes as the child grows.

Third, the use of logits at a point in time is indicated by limitations in the data source. In every year, the *NLS-Y* interview collects the start and end dates of jobs and gaps in employment of one week or more that were not covered by paid vacation or sick leave. Paid maternity leave is often, therefore, reported as time worked. Consequently, the data on exact dates of return to work after pregnancy are unreliable. However, maternity leave questions asked in 1983 indicate that almost all of the maternity leaves (82 percent) were completed within 3 months.

The sample includes first births to women who were not students or in the military at the time of the birth, and who had worked more than 20 hours per week at some time prior to the birth. The wage rate measure is the hourly wage at the last job held prior to the birth. All other covariates are measured as of the interview closest to the birth. Wages and income are expressed in 1987 dollars.

Our analysis includes several measures of child care costs. The presence of the grandmother or other females over age 10 in the household indicates access to low-cost or free child care from relatives. We calculated mean hourly earnings of child care workers from the 1980 Census to measure the variability of child care labor inputs by state and urban areas within states (Blau and Robins).

---

[1] We recognize that the logit specification imposes a strong independence assumption (Independence of Irrelevant Alternatives) on the residuals. We suspect that the decisions to use relative and nonrelative child care rather than to not work are correlated, which would violate that assumption. We will relax this assumption in later work. The standard errors are computed following H. White (1982), and are therefore robust to this potential specification error.

Data on the maximum child care tax credit (in 1987 dollars) and the marginal rate of child care credit measure the variation over time and geographical area in the net cost of child care. Between 1979 and 1981, the nominal federal child care tax credit was a maximum of $400 for a family with one child. The maximum credit grew to $480 in 1982. It remained constant at that level, although its real value declined with inflation. Beginning in 1983, the maximum credit and the marginal credit are decreasing functions of household income. We computed maximum and marginal credits for the household income obtained by summing the spouse's income plus the woman's income if she worked full time at her last wage. We also include data on state child care tax credits compiled by Blau (1989).

III. Results

The young women in the *NLS-Y* sample returned to work rapidly after the birth of their first child. More than one-third were at work in the first 3 months after delivery, and three-quarters had returned before their infant was 2-years old. Among those returning to work, more than one-half had relatives care for their infant. This was true both for women who returned in the first quarter after delivery (55 percent)[2] and for women who did not work in the first quarter, but were working by the time their child was 2-years old (56 percent). Nonrelative care accounted for about one-third of the child care arrangements—35 percent among women returning in the first quarter and 31 percent among women who returned to work between the first and eighth quarters. Thus, relatives appear to be nearly equally important for early and later returns to work. Child care centers and other out-of-home arrangements accounted for 8–11 percent of the child care.

TABLE 1—RETURN TO WORK FOLLOWING CHILDBIRTH

	In First 3 Months		In Months 3–24	
	(1)	(2)	(3)	(4)
Intercept	-10.720^{a}	-6.804^{c}	2.401	5.473
	(3.484)	(3.962)	(4.132)	(4.904)
Wage (log $)	0.586^{a}	0.674^{a}	0.262	0.597^{b}
	(0.193)	(0.226)	(0.195)	(0.269)
Husband's Income (1000 $)	-0.016^{b}	-0.029^{a}	-0.019^{b}	−0.016
	(0.007)	(0.010)	(0.008)	(0.011)
< H.S. Graduate	-0.676^{a}	-0.751^{a}	-0.483^{b}	-0.613^{b}
	(0.189)	(0.260)	(0.195)	(0.244)
Some College	0.136	0.686^{a}	0.136	0.569^{b}
	(0.177)	(0.197)	(0.254)	(0.274)
Birth Year	0.097^{b}	0.044	−0.017	−0.074
	(0.046)	(0.054)	(0.056)	(0.066)
Age at Birth	0.014	0.039	-0.083^{c}	-0.117^{b}
	(0.037)	(0.049)	(0.047)	(0.058)
Married	0.160	0.080	0.286	0.601^{b}
	(0.174)	(0.227)	(0.200)	(0.255)
Av. CC Wages in State ($)	0.062	0.088	0.089	0.249^{a}
	(0.065)	(0.080)	(0.087)	(0.097)
Mar. CC Tax Rate	−1.425	-4.238^{b}	−1.252	−0.624
	(1.162)	(1.715)	(1.275)	(1.653)
Max. CC Tax Credit	1.022^{a}	1.329^{a}	0.438	0.859
	(0.356)	(0.502)	(0.441)	(0.529)
Grandmother	0.337^{c}	-0.883^{a}	0.363	0.110
	(0.185)	(0.298)	(0.221)	(0.292)
Female >10	−0.103	−0.046	0.074	−0.022
	(0.108)	(0.149)	(0.116)	(0.156)
No. of Observations	1523		927	
Log Likelihood	−1226.55		−813.64	

Note: Multinomial Logit Estimates from *National Longitudinal Survey of Youth* data, 1979–86. Coefficients show effect relative to the excluded category, did not return to work. Standard errors are shown in parentheses.

Cols. 1 and 3 denote Relative Care and Cols. 2 and 4, Nonrelative Care.

$^{a}.01 < p$.

$^{b}.05 < p \leq .01$.

$^{c}.10 < p \leqq .05$.

The effects of the independent variables on the log odds of working and using care by relatives, and working and using care by nonrelatives relative to not returning to work are shown in Table 1. We hypothesized that women who had access to relatives to provide child care would return to work sooner after childbirth. For returns to work within the first 3 months (cols. 1 and 2), the presence of a grandmother in the home leads to a greater probability of working and using care by relatives, and a lower probability of working and using nonrelative care. The net effect on the overall probability of working (at the means of the other variables) is, however, not significantly different from zero. Married women and women living in a household with other females over age 10 were no more likely than other women to return to work after delivery.

---

[2] All percentages are weighted.

For returns to work after the first 3 months (cols. 3 and 4), the importance of relatives is attenuated. There is no significant effect of the "grandmother" on either type of child care use. Married women are more likely to work using nonrelative care if their husband's income is below $37,000. For husbands earning more than that, the negative effect of family income offsets the positive effect on labor supply of being married. For returning to work with relative care, the effect of being married is offset by greater family income if the husband earns more than $10,000.

The maximum value of the child care credit is positively related to returns to work within 3 months of delivery with both relative and nonrelative care. The marginal tax care credit has a negative effect on returning to work with market care, and does not affect working with care by relatives. Neither child care tax variable significantly affected returns to work after the first 3 months. Average child care wages are positively related to the probability of working in all four equations. However, the effect is significant only for women returning to work between 3 and 24 months after delivery, and using nonrelative care. Thus, the causality may be reversed, with higher wages for child care workers reflecting areas with a greater demand for care, or greater demand for female workers generally.

We also confirm that women with higher wages and more education return to work with either type of child care both soon after childbirth (3 months) and later (2 years). A woman with a higher-income spouse is slower to return to the labor market. Women who gave birth in later years were more likely to return to work when their infant was quite young (less than 3 months), and less likely to return to work when their child was between 3 and 24-months old.

IV. Discussion

The expected level of child care tax credits is positively related to the likelihood of a mother returning to work within 3 months of her first child's birth. Though positive, the effect of child care tax credits is diminished for later returns to work. This is consistent with the somewhat greater importance in the earlier period of the financial variables, such as wages and family income. The strong effects relating to the presence of the grandmother suggest that nonfinancial aspects of care (for example, knowing the caregiver) are also important for returns to work shortly after childbirth, particularly, for the choice of child care mode. Indeed, women report that care by a relative is their preferred type of child care for infants (Mason and Kohlthau).

These results suggest that the more generous tax treatment of child care expenses, along with improved market opportunities for women and poorer opportunities for their spouses, have tended to promote earlier return to work among new mothers during the 1980's. It is interesting to note that child care tax credits have their greatest effect on labor supply among mothers of very young children. It is this group that has experienced the greatest rates of increase in market work during the 1980s.

REFERENCES

Blau, David, M., "The Labor Market Effects of Child Care Subsidies," unpublished manuscript, University of North Carolina, 1989.

______ **and Robins, Philip K.,** "Fertility, Employment, and Child Care Costs," *Demography*, May 1989, *26*, 287–99.

Connelly, Rachel, "The Effect of Child Care Costs on Married Women's Labor Force Participation," presented at Population Association Meetings, Baltimore, March 1989.

Heckman, James J., "Effects of Child-Care Programs on Women's Work Effort," in T. W. Schultz, ed., *Economics of the Family*, Chicago: University of Chicago Press, 1974, 491–518.

Leibowitz, Arleen, Klerman, Jacob and Waite, Linda, "Women's Employment During Pregnancy and Following Birth," presented at the Econometric Society, December 1988.

Mason, Karen O. and Kohlthau, Karen, "Determinants of Child Care Ideals Among Mothers of Preschool-Aged Children," *Journal of Marriage and the Family*, August 1989, *51*, 593–603.

O'Connell, Martin and Bloom, David, "Juggling Jobs and Babies: America's Child Care Challenge," in *Population Trends and Public Policy*, No. 12, Population Reference Bureau, Washington, February 1987.

Robins, Philip K., "Federal Financing of Child Care: Alternative Approaches and Economic Implications," presented to the Child Care Action Campaign Conference, New York, May 1988.

White, H., "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, January 1982, *50*, 1–26.

U.S. Bureau of the Census, *Statistical Abstract of the United States: 1988*, 108th ed., Washington: USGPO 1987.

Intercohort Differences in Women's Labor Market Transitions

By M. ANNE HILL*

After remaining virtually constant during the post-World War II period, the ratio of women's earnings to men's rose from 62 to 70 percent between 1979 and 1987 (based on the median usual weekly earnings of full-time workers aged 16 years and over). One explanation for the rise in women's relative earnings is that there has been an intercohort increase in the continuity of women's labor force participation. A number of cross-sectional studies have found that gender differences in the level and continuity of work experience over the lifetime help to explain much of the wage gap (Jacob Mincer and Solomon Polachek, 1974, 1978; Mary Corcoran and Greg Duncan, 1979). And although female labor force participation rates have risen dramatically during the post-World War II period, work by June O'Neill (1985) and by James Smith and Michael Ward (1985) implies that this increase actually reduced the lifetime work experience of the average working woman.

This paper analyzes the extent to which the labor force behavior of recent cohorts of women has actually changed by comparing the early labor force experience of women who were between the ages 16 and 21 in 1968, with those between the ages 16 and 21 in 1979. Eight-year samples from the *National Longitudinal Surveys of Young Women and Youth* form the data base for this research. For comparability, the *NLS-Young Women* are followed from 1968 through 1975, while the *NLS-Youth* include observations from the period 1979 through 1986.

I. The Empirical Approach

This paper builds on recent work using longitudinal data to analyze labor supply behavior and the implications of labor force patterns for the wage gap.[1] Notable among them is John Donohue (1987) who used *NLS* data to estimate the duration of the first postschooling job for samples of young men and women from two 4-year periods, 1968–71 and 1979–82. His results indicate that, for the more recent cohort, the early labor force behavior of young women appears virtually identical to that of young men.

The objective here is to explain variation in the number of consecutive years a woman is working (i.e., the duration of a work spell) as well as consecutive years not working. Given the dynamic nature of female labor supply and the availability of event history information, I use continuous time methods to model the determinants of exits from work spells as well as from nonwork spells.

The *NLS* data include both completed spells, the duration of which is known, and *censored* spells, for which the end of the labor force spell is not yet observed. Maximum likelihood estimation makes full use of the information for completed and censored work and nonwork spells. I use information on all spells for all women, which implies multiple spells for some women.[2] The analysis includes those persons who eventually leave the sample if we observe at least 3 years data for them; their work and nonwork spells are treated as being right-censored.

Spell lengths are measured in years; if an individual reports herself to be employed for

---

*Department of Economics, Queens College, and Center for the Study of Business and Government, Baruch College, CUNY, 17 Lexington Avenue, New York, NY 10010. This research is part of a larger project conducted jointly with June O'Neill at the Center for Business and Government and funded by BLS grant no. E-9-M-8-0092. Hengzhong Liu and Partha Deb provided excellent programming assistance.

[1] See also work on dynamic labor supply by C. J. Flinn and James Heckman (1982, 1983).

[2] See Kei-Mu Li et al. (1987), Flinn-Heckman (1982, 1983), and Heckman and B. Singer (1986) for descriptions of the multiple-spell hazard rate model and estimation.

at least 6 months of a given year, that year is counted as a work spell year. Otherwise the year is defined as a nonwork spell year.[3]

The explanatory variables include both those that are fixed for a given spell and those that vary with time as the spell progresses. Fixed variables consist of years of experience, age, and years of schooling, each measured at the beginning of the spell. Time-varying variables include a dummy variable equal to one if the woman is enrolled in school during the spell year, two time-varying residential dummies that capture southern and SMSA residence, and vectors of dichotomous marital status and fertility variables (included to capture changes in household composition). Women who remained unmarried in a spell year define the reference category for marital status, while women who have no children are the reference fertility group. The model also includes spell duration and its square.

II. Empirical Results

Given that labor force behavior may differ by race, models are estimated separately for nonblack and black women.[4] The most striking intercohort difference in spell characteristics is that the mean duration of *censored* work spells increases by more than 1 year for the younger cohort of women, rising from 4.0 to 5.2 years (of a potential 8) for nonblack women and from 3.4 to 4.5 years for black women. Most other spell durations remain quite similar for the two cohorts, except that censored nonwork spells have lengthened for the younger group of black women, with mean duration rising from 4.3 to 4.9 years.

TABLE 1—WORK (W) AND NONWORK (NW) SPELL HAZARD ESTIMATION, NONBLACK WOMEN

	NLS-Young Women ($N=1761$)		*NLS-Youth* ($N=2716$)	
	(W)	(NW)	(W)	(NW)
INTERCEPT	-3.579^{c}	-5.178^{c}	-3.680^{c}	-5.053^{c}
SOUTH	0.007	−0.007	−0.007	0.063
CITY CENTER	−0.055	−0.028	−0.046	-0.142^{b}
AGE/100	6.836^{c}	0.412	4.072^{b}	-3.539^{b}
EXPERIENCE/100	-6.104^{a}	9.806^{c}	-11.512^{c}	9.173^{c}
SCHOOLING/100	-15.822^{c}	15.683^{c}	-15.768^{c}	12.558^{c}
ENROLLED	0.593^{c}	-0.218^{c}	0.287^{c}	0.274^{c}
SEPARATED	−0.111	−0.476	0.206	-0.498^{b}
BECAME MARRIED	0.148	-0.474^{c}	0.110	-0.304^{c}
REMAINED MARRIED	0.089	-0.874^{c}	0.108	-0.408^{c}
FIRST BIRTH	0.788^{c}	-0.286^{b}	0.830^{c}	-0.384^{c}
HIGHER-ORDER BIRTH	−0.219	−0.245	0.142	-0.258^{b}
NO BIRTHS, KIDS < 6	0.209^{a}	-0.640^{c}	0.518^{c}	-0.485^{c}
ALL CHILDREN ≥ 6	-1.932^{c}	−0.505	-0.362^{a}	-0.456^{b}
DURATION/10	6.337^{c}	10.070^{c}	5.276^{c}	8.858^{c}
DURATION^2*/100*	-20.155^{c}	-27.229^{c}	-19.295^{c}	-26.875^{c}
Negative Log Likelihood	202.1		749.5	

[a] $0.05 < \alpha \leq 0.10$.
[b] $0.01 < \alpha \leq 0.05$.
[c] $\alpha \leq 0.01$.

TABLE 2—WORK (W) AND NONWORK (NW) SPELL HAZARD ESTIMATION, BLACK WOMEN

	NLS-Young Women ($N=718$)		*NLS-Youth* ($N=1017$)	
	(W)	(NW)	(W)	(NW)
INTERCEPT	-7.007^{c}	-3.164^{c}	-6.173^{c}	-2.997^{c}
SOUTH	−0.137	0.015	-0.365^{c}	0.178^{b}
CITY CENTER	0.027	0.065	0.072	−0.102
AGE/100	−2.631	0.748	4.248^{a}	-5.608^{b}
EXPERIENCE/100	−5.495	16.548^{c}	-19.857^{c}	19.122^{c}
SCHOOLING/100	-13.234^{c}	16.895^{c}	-19.028^{c}	21.129^{c}
ENROLLED	0.296^{b}	−0.100	0.376^{c}	0.457^{c}
SEPARATED	−0.100	−0.654	−0.142	0.070
BECAME MARRIED	−0.266	−0.107	0.099	−0.050
REMAINED MARRIED	0.112	-0.262^{b}	-0.371^{b}	−0.002
FIRST BIRTH	0.499^{b}	−0.033	0.365	−0.139
HIGHER-ORDER BIRTH	0.387	−0.051	0.180	−0.021
NO BIRTH, KIDS < 6	−0.077	-0.520^{c}	0.400^{c}	-0.547^{c}
ALL CHILDREN ≥ 6	−1.218	-1.180^{b}	0.112	-0.738^{c}
DURATION/10	8.442^{c}	6.689^{c}	7.783^{c}	6.949^{c}
DURATION^2*/100*	-35.292^{c}	-16.084^{c}	-30.464^{c}	-16.823^{c}
Negative Log Likelihood	75.6		297.0	

[a] $0.05 < \alpha \leq 0.10$.
[b] $0.01 < \alpha \leq 0.05$.
[c] $\alpha \leq 0.01$.

Tables 1 and 2 present the estimates of the multiple spell hazard rate models for nonblack and for black women, respectively.

---

[3] The *NLS-Young Women*'s sample began skipping years in the late 1970s so a full sequence of annual observations is not available. However, through the use of retrospective questions, it is possible to fill in important information for the missing years. Given the nature of the retrospective questions, I define a work year as one in which the respondent worked at least 6 months to link the retrospective and prospective data. Unfortunately, this definition combines together the states "unemployment" and "out of the labor force" as "nonwork."

[4] To adjust for the oversampling of Hispanics in the *NLS-Youth*, I drew a random sample of all Hispanic women such that my sample proportion corresponds with the age and sex-specific population proportion as reported in the 1980 Census.

Considering first the estimated risk of leaving a work spell, southern residence is significantly associated with longer work duration among younger black women, while SMSA residence is not statistically significant. The conditional probability of leaving work increases with age except among the older group of black women, for whom this effect is not statistically significant.

Prior labor force experience and schooling reduce the conditional probability of leaving a work spell. For both nonblack and black women, the effect of experience has strengthened over time, dramatically so for black women. Being enrolled in school increases the likelihood of exiting a work spell. Surprisingly, changes in marital status have a statistically significant effect on work spells only for the younger cohort of black women, for whom remaining married increases the duration of a work spell. Relative to a childless woman, one who gives birth to her first child during a work-spell year is much more likely to end the work spell. This effect is weaker for black women, but differs little by cohort. Bearing a second or higher-order child while working has no statistically significant effect on leaving work, yet the presence of children under age 6 increases the probability of ending a work spell (except for the older cohort of black women). Generally, the presence of children who are older than 6 prolongs work spells. Finally, after roughly 3 months in a work spell, the effect of duration becomes negative. This supports the conjecture that cumulative experience *within* a spell raises otherwise unmeasured returns to continued work participation, and reduces the probability of leaving work.

Turning to the estimates for nonwork spells, residing in the South exerts a statistically significant effect to reduce the length of a nonwork spell only for the younger group of black women, while SMSA residence lengthens the nonwork spells of younger nonblack women. Being older significantly extends nonwork spells only for the younger cohort. The risk of leaving a spell of nonwork increases with experience and schooling. Again, these effects are stronger for the younger cohort of black women. The consequence of school enrollment appears to change over time for both racial groups. Among the older cohort, school enrollment lengthens a nonwork spell, while the opposite holds for younger women. The impact of marital status on the nonwork spells of white women has diminished over time. Relative to remaining single, separating, getting married and staying married inhibit labor market entry for white women. Only among the older cohort of black women does marital status bear significantly on nonwork spells. For each group, childbearing reduces the likelihood of leaving a nonwork spell. Among white women, these variables maintain roughly the same order of magnitude for the more recent cohort. For black women, a first or higher-order birth during a nonwork spell year has an insignificant effect on the likelihood of leaving the spell. Only the presence of children appears to matter, with the effect of older children being slightly weaker for the younger cohort.

What do these results inform us about intercohort differences in labor force transitions? Considering the dependent variable alone, the length of censored work spells has risen nearly 1 year (of a potential 8) for both white and black women. The human capital variables, in particular the level of labor market experience and schooling, both increase the duration of work spells and hasten an exit from a nonwork spell. These effects appear stronger for the younger cohort of women, especially black women. Yet the intercohort differences in responses to demographic variables appear mixed. Relative to being single, all marital states deter exists from nonwork spells; and these effects are weaker among the younger cohort of women. Yet, the labor supply effects of childbearing appear to have diminished only slightly among the younger cohort.

REFERENCES

Corcoran, Mary and Duncan, Greg, "Work History, Labor Force Attachment, and Earnings Differences Between the Races and Sexes," *Journal of Human Resources*, Winter 1979, *14*, 3–20.

Donohue, John J., "The Changing Relative Hazard Rates of Young Male and Female Workers," unpublished paper, Northwestern University School of Law, American

Bar Foundation, September 1987.

Flinn, C. J. and Heckman, J. J., "Models for the Analysis of Labor Force Dynamics," *Advances in Econometrics*, Vol. 1, Greenwich: JAI Press, 1982, 35–95.

______ and ______, "The Likelihood Function for the Multistate-Multiepisode Models in Models for the Analysis of Labor Force Dyamics," *Advances in Econometrics,* Vol. 2, Greenwich: JAI Press, 1983, 225–31.

Heckman, J. J. and Singer, B., "Econometric Analysis of Longitudinal Data," in Zvi Grilliches and Michael D. Intrilligator, eds., *The Handbook of Econometrics*, Vol. III, Amsterdam: Elsevier, 1986.

Li, Kei-Mu, Honore, Bo and Walker, James, *CTM: A Program for the Estimation and Testing of Continuous Time Multi-State Multi-Spell Models*, Chicago: University of Chicago Press, 1987.

Mincer, Jacob and Polachek, Solomon, "Family Investments in Human Capital," *Journal of Political Economy*, March/April 1974, *82*, S76–S108.

______ and ______, "Women's Earnings Reexamined," *Journal of Human Resources*, Winter 1978, *13*, 188–33.

O'Neill, June, "The Trend in the Male-Female Wage Gap in the United States," *Journal of Labor Economics*, January 1985, Suppl., *3*, S91–S116.

Smith, James P. and Ward, Michael P., "Time-Series Growth in the Female Labor Force," *Journal of Labor Economics*, January 1985, Suppl., *3*, S59–S90.

Gender Differences in Wages and Job Turnover Among Continuously Employed Workers

By AUDREY LIGHT AND MANUELITA URETA*

The labor force participation rate of American women has risen steadily since World War II. This may reflect an increase in the number of women entering the labor force to pursue "spotty" careers, a rise in the number of continuously employed women, or both. The distinction is important, for women in the latter group are likely to invest more intensively in human capital during their early careers, and they may also engage in "job shopping." As a result, they should receive more wage growth and end up with longer jobs than their irregularly employed counterparts. In fact, these women may have careers that are indistinguishable from those of men.

Of course, women who work continuously could also behave differently than men. Although family obligations do not dissuade these women from working, they may influence their choice of occupation and hours. For example, women may favor 9 to 5 jobs, or jobs that are located close to their homes. Furthermore, if employers believe that women are "quitters," then they will be reluctant to share in the cost of on-the-job training. Even if the increase in female labor force participation has produced a growing pool of continuously employed women, the implications for gender differences in earnings and job mobility are unclear.

The notion that a given participation rate is consistent with a range of underlying behavioral patterns is not new. Yoram Ben-Porath (1973) observes that a homogeneous population of sporadic participants and a heterogeneous population of continuous workers and nonworkers can produce the same average participation rate. James Heckman and Robert Willis (1977) examine a sample of married women and find that the latter scenario appears to hold. In fact, they find that most women have participation probabilities for a given year that are close to either zero or one. However, since the size and composition of the female labor force is changing rapidly, a reinvestigation of this issue is warranted.

The purpose of this paper is to determine whether a significant number of women work continuously during their early careers, which women are likely to do so, and how these women compare to men in terms of their interfirm mobility and earnings. The data are from the young men and young women cohorts of the *National Longitudinal Surveys of Labor Market Experience* (*NLS*). All respondents were between ages 14 and 24 when the surveys began. Although the women were followed from 1968 to 1985, and the men from 1966 to 1981, we restrict ourselves to the postschool portion of each respondent's panel. Our sample consists of 4,490 women and 4,600 men.[1]

I. Labor Force Participation Patterns

We begin by computing the fraction of each worker's time that is spent working. Because it is not possible to account for each respondent's activities during the entire survey period, we calculate the number of weeks worked as a fraction of time that we are able to account for. The relatively older and/or less educated individuals in the sample are observed over a longer period of time than are the others. To eliminate any biases this may create, we compute the fraction of time

---

*State University of New York, Stony Brook, NY 11794, and The Claremont Graduate School, Claremont, CA 91711, respectively.

[1] See our earlier paper (1989) for a complete description of the samples and for details on the hazard models discussed in Section II.

spent working between the ages of 24 and 30.

Table 1 shows the distribution of this "work continuity" measure for women. Because blacks and whites prove to differ dramatically in their patterns of labor force participation, we divide the sample by race in this table and in most of our subsequent analysis. Looking across the top two rows of Table 1, we see that the distribution for blacks looks very similar to the distribution for whites (in fact, the difference between these distributions is statistically insignificant): 89 percent of black women and 87 percent of whites spend more than 10 percent of their time working, while 28 percent of blacks and 25 percent of whites spend more than 90 percent of their time working.

Controlling for schooling reveals that the first two rows of Table 1 mask important differences between black and white women. Within race, the distribution shifts to the right as schooling levels increase. Only 11 percent of white high school dropouts work more than 90 percent of the time, while almost one-third of the college graduates attain this level of work continuity. For black women, the corresponding numbers are 19 and 45 percent. These numbers also reveal that, within most schooling levels, the distribution for blacks lies to the right of the distribution for whites.[2] This domination is particularly pronounced among college graduates. Since a much larger proportion of whites than blacks attains this top schooling level, we conclude that the similarity of the overall distributions (rows 1–2) simply reflects the educational discrepancy across races.

TABLE 1—PERCENT OF WOMEN WHO WORK MORE THAN *X* PERCENT OF THE TIME FROM AGE 24 TO 30

	10	30	50	70	90	Sample Size
Whites	87	74	60	46	25	3167
Blacks	89	76	63	48	28	1323
By Schooling Level at Age 24						
Whites						
0–11 years	78	58	42	25	11	318
12 Years	83	68	54	39	22	1217
16+ years	95	85	74	59	32	469
Blacks						
0–11 years	78	62	46	33	19	285
12 years	90	79	67	50	29	419
16+ years	100	95	88	76	45	104
By Birth Cohort						
Whites						
1944–47	83	66	54	39	18	764
1951–54	89	76	62	48	28	929
Blacks						
1944–47	88	76	64	47	26	231
1951–54	89	78	64	51	32	415
By Birth Cohort and Schooling Level						
Whites						
1944–47						
0–11 years	79	64	54	32	15	78
12 years	79	61	49	35	15	391
16+ years	93	78	66	49	24	152
1951–54						
0–11 years	78	56	39	28	11	140
12 years	87	73	59	44	27	447
16+ years	99	91	81	67	43	187
Blacks						
1944–47						
0–11 years	81	64	48	34	20	64
12 years	88	76	62	44	26	111
16+ years	100	73	73	60	47	15
1951–54						
0–11 years	80	66	47	33	20	132
12 years	91	80	65	51	32	179
16+ years	100	95	86	77	49	57

In order to control for birth cohort, we define an "early" cohort made up of women born in 1944–47, and a "late" cohort consisting of women born in 1951–54. For whites, we find that the late cohort's distribution lies well to the right of that for the early cohort. Only 18 percent of the older women work more than 90 percent of the time, while 28 percent of the younger ones do. Given that only 7 years separate the two cohorts, this is a very dramatic shift toward continuous employment. Equally remarkable is the finding that the distribution for blacks is virtually unchanged over time.

---

[2] Kolmogorov-Smirnov statistics reveal that this domination is statistically significant at a 99 percent confidence level only for schooling levels 12 and 16+. The other distributional dominances that are statistically significant are: (a) within each race, 12 years of schooling over 0–11 years and 16+ years over 12 years, (b) whites born in 1951–52 over whites born in 1944–45, (c) white high school graduates born in 1951–52 over white high school graduates born in 1944–45, and (d) white college graduates born in 1951–52 over white college graduates born in 1944–45.

Table 1 also shows the distributions after controlling for race, birth cohort, and schooling level. Among high school dropouts, the distribution for blacks is invariant over time, while the distribution for whites actually shifts to the left (although not significantly). Clearly, the increase over time in the proportion of women who work "continuously" is coming strictly from well-educated women, especially whites. Within every cohort-schooling group, however, we can conclude that a black woman is likely to spend a larger proportion of her time working than is a white woman.

II. Job Turnover

Based on the figures in Table 1, we can infer that the postwar influx of women into the labor force consisted of a significant number of women who work (more or less) continuously. Since career interruptions are viewed as the cause of high turnover rates among women, it seems natural to investigate the mobility of women who clearly are *not* quitting their jobs to leave the labor force. In particular, we wish to learn whether these women use the early parts of their careers to lock into good jobs, as men typically do.

To address this issue, we estimate proportional hazard models for all job separations among continuously employed workers. Since the median man in our sample works for 88 percent of the time that we are able to account for, we use this number as the cutoff for continuously employed workers. That is, we define men and women to be continuously employed if they work for a larger fraction of their time than the median man. Our sample of continuously employed workers includes 2,343 men and 1,142 women.

We estimate the hazard model separately for four gender-cohort groups and include race as a covariate. We also control for marital status, children, schooling, prior experience, current wages, hours, union status, industry, year, and whether the worker resides in an SMSA and the South. These estimates are used to compute the implied probabilities that the "modal" worker will leave his or her job in the next 6 months, conditional on various levels of current tenure. The modal worker is married, lives in an SMSA, is a high school graduate, and is deemed to be a "stayer" for unobserved reasons.[3] The probabilities (shown in Table 2) are evaluated at the cohort mean values of the hourly wage, unemployment rate, and prior experience; the probabilities are evaluated in 1970–71 for the early cohort and in 1976–77 for the late cohort.

TABLE 2—CONDITIONAL PROBABILITY OF JOB SEPARATION IN THE NEXT 6 MONTHS FOR 24–30 YEAR OLD CONTINUOUSLY EMPLOYED WORKERS

Months of Tenure	Whites		Blacks	
	Women	Men	Women	Men
Born 1944–46				
0	.103	.167	.098	.190
12	.082	.132	.078	.150
24	.064	.104	.061	.118
36	.051	.081	.048	.093
48	.040	.063	.037	.073
60	.031	.049	.029	.057
Born 1950–52				
0	.199	.188	.188	.157
12	.173	.159	.163	.132
24	.150	.134	.141	.111
36	.129	.112	.122	.093
48	.112	.094	.105	.078
60	.096	.079	.090	.065

Table 2 reveals that, among workers in the earlier cohort, the probability of separation is lower for women than for men regardless of the current level of tenure. White women have a 10.3 percent chance of leaving their jobs in the first 6 months, while men have a 16.7 percent chance. If these workers' jobs last for 60 months, the probability of leaving in the next 6 months falls to 3.1 for women and 4.9 for men. The difference between men and women is even greater for blacks. When we focus on the later cohort, however, we see that women are more likely than men to leave their jobs in the next 6 months, regardless of current tenure and race.

---

[3] The hazard model estimates (which are available from the authors) reveal that 35 percent of the women born in 1944–46 and 95 percent of the women born in 1950–52 are stayers. For the men, the corresponding numbers are 52 and 57 percent.

Of course, high turnover rates are not "bad" if they indicate that workers are moving to better jobs. While men and women differ in their separation probabilities, they exhibit virtually the same degree of (negative) duration dependence. That is, the rate at which their separation probabilities decline with tenure are the same, regardless of cohort and race. If this means that workers are finding and locking into good jobs, then it is good news for men and women alike.

III. Wage Differentials

To see how the earnings of these continuously employed men and women compare, we estimate hourly wage equations for each race-cohort group. Gender is included as a regressor, as are interactions between gender and schooling, tenure and experience. We also control for marital status, children, tenure, and experience (and their quadratics), part-time status, union status, time, and whether the worker resides in an SMSA and in the South. Table 3 shows the implied ratios of female-to-male wages for workers with no job tenure and 4 years of work experience.

TABLE 3—RATIOS OF FEMALE-TO-MALE STARTING WAGES FOR 24–30 YEAR OLD WORKERS

	Schooling Level			
	0–11	12	13–15	16+
Continuously employed				
Whites				
Born 1944–46	0.58	0.66	0.75	0.72
Born 1950–52	0.82	0.77	0.82	0.84
Blacks				
Born 1944–46	0.72	0.73	0.70	0.85
Born 1950–52	1.02[a]	0.86	0.84	0.80
All workers				
Whites				
Born 1944–46	0.55	0.59	0.69	0.72
Born 1950–52	0.63	0.68	0.76	0.77
Blacks				
Born 1944–46	0.62	0.61	0.72	0.90[a]
Born 1950–52	077	0.77	0.82	0.93[a]

[a] Ratio is *not* significantly different from one at 95 percent confidence level.

Table 3 reveals that the wage differential between continuously employed men and women is closing over time for almost every race-schooling group. For example, the typical white woman in the 1944–46 cohort with 0–11 years of schooling makes 58 cents to her male counterpart's $1, while a woman born in 1950–52 makes 82 cents. Among blacks in the same schooling group, the ratio increases from 0.72 to 1. The only group that does not show improvement over time is blacks in the top schooling group, where the wage ratio falls from 0.85 to 0.80.

Table 3 also presents comparable statistics for all workers, regardless of their degree of employment continuity. With few exceptions, these ratios are lower than the corresponding numbers in the top panel. Among black women, those with 13–15 years of schooling in the early cohort and college graduates in both cohorts do better relative to men if they are drawn from the full sample than if they are continuously employed. Among white women, college graduates in the early cohort earn 72 percent as much as men regardless of which sample they are drawn from. In every other group, however, women do better relative to men if they are continuously employed.

Although the wage gap is closing over time for the full sample of workers, the gains are less dramatic than what we see in the top panel. Among white college graduates, for example, the wage ratio for the full sample increases from 0.72 to 0.77, which is far less than the gain seen among the continuous workers. In fact, the only group where continuously employed workers show smaller gains than the full sample is black college graduates (which is the smallest sample examined). Whereas the wage ratio for continuously employed black college graduates has fallen, the corresponding ratio in the bottom panel is one for both birth cohorts.

IV. Conclusions

Continuous employment is by no means the norm among young women, but it appears to be a growing trend. Barely more than one-fourth of the women in our sample spend more than 90 percent of their time working when they are 24 to 30 years old, but roughly 45 percent of the younger college graduates attain this level of labor force attachment. Because schooling levels of women are rising very rapidly, it is likely that the fraction of women who remain firmly attached to the labor force will increase even more in the future.

What effect does this increased labor force attachment have on gender differences in earnings and interfirm mobility? In comparing the job turnover behavior of continuously employed men and women, we find that both genders exhibit identical degrees of negative duration dependence. While women born in 1944–46 are less likely than men to leave their jobs (regardless of race, education, and current tenure), the opposite is true for a cohort born just 6 years later. In comparing starting wages of men and women, we find that the wage gap is less pronounced among continuously employed workers than among the full sample in almost every race-cohort-schooling group, and the gap is narrowing far more rapidly among the continuously employed.

While these comparisons reveal genuine improvements in the relative labor market status of women, we plan to enhance the analysis in several ways. First, we hope to learn whether women fare equally well as their jobs age and as they age beyond 30. We also wish to learn why black women have not increased the continuity of their employment over time, as whites have. In addition, we plan to compare trends in the educational and occupational attainment of continuously employed men and women.

REFERENCES

Ben-Porath, Yoram, "Labor Force Participation Rates and the Supply of Labor," *Journal of Political Economy*, May/June 1973, *81*, 697–704.

Heckman, James J. and Willis, Robert J., "A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *Journal of Political Economy*, February 1977, *85*, 27–58.

Light, Audrey and Ureta, Manuelita, "Panel Estimates of Male and Female Turnover Behavior: Can Female Non-Quitters Be Identified?" SUNY-Stony Brook Research Paper No. 328, June 1989.

Earnings Inequality and Job Turnover

By JONATHAN S. LEONARD AND LOUIS JACOBSON*

Inequality has increased in the United States since 1970 (Barry Bluestone, 1989, and Bluestone and Bennett Harrison, 1988). As the divergence in wages between skilled and unskilled workers increased, real wages for the median worker declined (Chinui Juhn et al, 1989). Neither of these troubling trends is well understood. Shifts in the supply of labor are unlikely to explain the increasing returns to education and experience because the labor force was becoming older and better educated (John Bound and George Johnson, 1989). This has led to a number of studies of labor demand shifts as a possible explanation for increasing wage variance.

The dramatic appreciation of the dollar during the early 1980s accompanied increasing trade deficits and industrial restructuring. The increasing openness of the U.S. economy to foreign competition, along with large trade deficits, are seen by Kevin Murphy and Finis Welch (1988) as the main cause of recent changes in the wage structure. In Lawrence Katz and Lawrence Summers (1988), these same forces are argued to have reduced rents formerly received by workers in trade-sensitive sectors. While foreign trade shifts did contribute to an acceleration during the 1980s of the ongoing decline in manufacturing and mining, this was offset by increases in construction and decreases in the public sector (Bound-Johnson). With no compelling evidence in support of the most obvious sources of demand or supply shifts, a number of researchers infer the existence of technological change favoring skilled workers (Bound-Johnson; Juhn et al.). This is consistent with evidence that the fastest growing enterprises within industry during the 1970s and 1980s were those with high proportions of skilled workers in technical and professional occupations (Leonard, 1986; 1988).

This paper uses earnings data from the State of Pennsylvania to examine the nature and sources of changes in the earnings distribution between 1973 and 1988. Pennsylvania is home to the steel industry and to related durable manufacturing industries that were hit hard during the early 1980s. This state serves as a test case for the proposition that the industrial restructuring of the early 1980s led to the shrinking middle class.

Many past studies have not been able to focus directly on labor market shifts because they analyzed income rather than earnings. Here we analyze earnings in a 0.6 percent sample of all Pennsylvania workers covered by Unemployment Insurance (UI). The data are derived from state UI tax records (see Jacobson, 1989). Because of changes in coverage, public sector workers are excluded here. Employees are classified into the industry of the employer contributing most of their earnings during a year.

Earnings are defined as total annual earnings in all employment covered by the Pennsylvania UI tax system. Employees are sampled as long as they report any positive earnings during the year. In particular, the unemployed or labor force nonparticipants with some earnings during the year are sampled. The annual earnings of employees switching jobs are counted once. Real earnings are obtained by dividing by the fixed-weight personal consumption expenditure price index (1987 = 100). The results of inequality studies are sensitive to choice of sample, period, income measure, deflator, and inequality index (McKinley Blackburn

---

[†] *Discussants*: Robert Topel, University of Chicago; Robert Moffitt, Brown University; Steven Davis, University of Chicago.

*Haas School of Business, 350 Barrows Hall, University of California, Berkely CA 94720 and NBER; and W. E. Upjohn Institute for Employment Research, Kalamazoo, MI, respectively.

et al., 1989; Michael Horrigan and Stephen Haugen 1988). In particular, evidence of increased earnings dispersion is stronger for full-time full-year workers than for all workers (Bluestone-Harrison; Marvin Kosters and Murray Ross 1987).

I. Industrial Restructuring

Pennsylvania represents a remarkable case of industrial restructuring. Between 1974 and 1987, nearly one-third of all manufacturing jobs were eliminated. Manufacturing's share of employment fell from 36 to 25 percent. Over the same period, total employment increased by nearly 20 percent, primarily due to increases in producer, social, and personal services, and in trade. This is a substantial shift of employment from high-wage to low-wage sectors. For example, on average, 46 percent of those employed in durable manufacturing earn more than $25,000 annually, but fewer than 22 percent of employees in producer services earn this much.

Despite this shift of employment from high- to low-wage sectors, the distribution of earnings did not change greatly. Between 1974 and 1987, the percentage of employees earning more that $30,000 increased from 12 to 13 percent, those earning from $20 to $30 thousand decreased from 18 to 16 percent, those earning from $10 to $20 thousand remained at 25 percent, and those earning less than $5,000 decreased slightly from 30 to 28 percent. The coefficient of variation (CV) of earnings increased slowly over these years from .91 to .98.

The standard deviation of the logarithm of real earnings declined slightly from 1.47 in 1975 to 1.45 in 1987, largely because of a trend increase in the proportion of workers employed for 4 quarters. Among 4-quarter workers, this measure increased from .76 to .82. This is in turn largely due to an increase in the between industry component of variation as average wages increased faster inside than outside manufacturing. Adding to this increase were the increased wage variance within manufacturing and the employment shift out of the relatively low variance manufacturing sector.

The overall wage distribution did not greatly change despite the employment shifts to low-wage sectors, because wages were increasing within nearly all sectors. The argument that rents declined in high wage industries under trade pressure must confront the fact that real wages increased in these sectors. For example, in durable manufacturing, the proportion of workers earning more than $25,000 increased from 35 percent in 1974 to 38 percent in 1987. In nondurable manufacturing, this proportion increased from 15 to 24 percent. These real wage increases occur not just in spite of employment losses, but partly because of these losses. The closure of high wage plants and moderation of wage increases is more than offset by the layoff of low tenure workers with relatively low wages within plants. Within tenure groups, real wage gains are attenuated or absent in manufacturing, and fall during the 1982 recession. At the same time, real wages also increase in the growing low-wage service sectors. In producer services, the percentage of workers earning more than $25,000 increased from 21 percent in 1974 to 26 percent in 1987. The aggregate wage distribution hides a number of complex changes. The shift from high wage manufacturing to low wage service and trade employment is offset by increases in wages within nearly all sectors.

II. The Active Margin

The active margin of the labor market should be a more sensitive indicator of shifts in the earnings distribution. This includes new employees, new jobs, and new employers who should be less subject to institutional rigidities.

New employees are those who have cumulated less than 4 quarters of employment with their primary employer by the end of the year. Between 1979 and 1987, the CV of new hires' earnings increased slightly from 1.34 to 1.49. This group is sensitive to changes in the duration of employment spells. Among those with between one and two years tenure, the CV of earnings increased from .85 to .90 over the same period. The earnings distribution of recent hires is less dispersed than that of the stock of employees, but this dispersion has increased slightly over time.

New employees in a given firm may replace previous workers in an ongoing job, and may not be new entrants to the labor market. An alternative way of slicing the active margin of the labor market is to examine earnings in growing enterprises. In growing enterprises, 30 percent of the workers earned less than $5,000 in 1976, while 19 percent earned more than $25,000. By 1986, these proportions had changed to 25 and 21 percent. By this measure the wage distribution of new jobs differs from that of old jobs. The wage distribution shifted upward in growing enterprises, but did not become substantially more dispersed.

Low-wage enterprises appear to grow more than do high-wage enterprises, absent other controls. Wages are higher in shrinking than in expanding enterprises. In 1987, 27 percent of the workers in shrinking enterprises earned more than $25,000. Part of this reflects industry differences. Part may also reflect age and size effects.

A third slice of the active margin of the labor market is new enterprises. In any given year, from 3 to 4 percent of all employment is in new enterprises. Enterprise age and wage are strongly correlated. Indeed, part of what has previously been identified as a size-wage effect may be due to an age effect: large enterprises are older, and their wage levels may have racheted up over time. In new enterprises, only about 8 percent of the workers earn more than $25,000. More than 20 percent of the workers in enterprises five or more years old earn this much. The entire wage distribution shifts upward as new enterprises age.

New enterprises cannot, however, account for much of a shift in earnings distribution. There is no strong trend shifting the earnings distribution of successive cohorts of entering enterprises. While wages are lower in new enterprises, they are not more dispersed. More importantly, the inflow of new jobs through new enterprises is too modest to have much effect over the wage distribution.

The outflow rate through enterprise closure is slightly larger, accounting for 2 to 6 percent of all employment. This rate often falls in a recession because most closures occur among the newborn, and birth rates often fall in a recession. Because new enterprises have higher mortality rates, the wage distribution in closed plants is shifted downwards toward that of new plants. About 32 percent of the workers in these enterprises earned less than $5,000 prior to closing. The distribution is more dispersed than in new enterprises: about 18 percent earned more than $25,000. In enterprises that continue operations, these percentages are 30 and 19 percent, and in new enterprises, they are 36 and 15 percent. Again, because the rate of employment loss through enterprise closure is modest and the wage distribution so similar to that in ongoing enterprises, enterprise closure does not directly shift the wage distribution by an appreciable amount. However, these salient events may have substantial indirect effects on wage bargaining.

Taken together, these different cuts from the active margin of new employees, new jobs in growing enterprises and from new enterprises tell a consistent story: the earnings distribution in Pennsylvania has shifted, but not greatly.

III. The Instability of Jobs

If one thinks of the U.S. economy as made up of a large proportion of stable jobs, bouts of industrial restructuring as in the early 1980s are a massive source of unusual job displacement. Indeed, an unusually high proportion of high-tenure and high-wage workers were laid off in 1982 and 1983, with lasting effects. Given that most of these displaced workers returned to employment, it may seem surprising that they did so with so little resulting change in the earnings distribution.

The U.S. labor market is actually far more fluid than this story allows for. Even in a normal year, the economy adjusts to large gross flows. In an average year during the early 1980s, one in every eight jobs was newly created, while one in every nine was destroyed (Leonard, 1987). Job creation is defined here as the proportion of net jobs added at growing enterprises.

For the manufacturing sector, S. J. Davis and John Haltiwanger (1989) report that in an average quarter during the early 1980s, 6 percent of all jobs disappeared and 5 percent were created. From year to year Davis and Haltiwanger report 8 percent job gain and 13 percent job loss rates. All such measures reflect a length bias: the job turnover rates reported by Timothy Dunne et al. (1989) on the basis of the *Census of Manufactures* (every 5 years) are correspondingly lower. During an average year, Leonard (1987) reports that 45 percent of all establishments grow by an average of 30 percent, while 47 percent of the establishments shrink by an average of 21 percent. Employment levels are anything but stable at the average establishment. These establishment job turnover rates are largely idiosyncratic, with only weak industry, region, or cyclical components. This within-industry fluctuation may account for a substantial portion of what is often thought of as the "natural" rate of unemployment. Both Leonard (1987) and Davis-Haltiwanger present sketchy evidence that hints that the instability of jobs increased during the 1980s. Complementary evidence indicates that the tenure distribution has shifted downward despite an aging workforce (Daniel Mitchell, 1989).

In Pennsylvania, 28 percent of all employees had accumulated less than one year's tenure with their primary employer during an average year from 1979 to 1987. Only 33 percent had more than four years tenure (uncompleted spell). In other words, more than one-quarter of all employees start employment spells in the year of observation. As the job turnover studies indicated, large gross flow rates characterize the economy. Enterprise turnover occurs at rates too low and in enterprises too small to account for much of this.

In this short time-series, no strong trend toward shorter tenture is observed. The tenure distribution is volatile. It shifts upward with the 1982 recession as hiring rates fall and low-tenure workers are laid off.

IV. Shifts in Earnings Mobility

Procyclical real wages due to shifting labor demand schedules are predicted both by real business cycle models and by Keynesian models. Given an upward-sloping short-run aggregate labor supply function, real wages are bid up as labor demand shifts upward. While aggregated data often show countercyclical or noncyclical wage changes, the predicted procyclical pattern is observed in longitudinal analyses of individual workers (Robert Barksy and Gary Solon, 1989). This procyclical variation is obscured by cyclical hiring and promotion rates in aggregated data.

We observe procyclical real earnings movements at the individual level. For each pair of successive years from 1974 to 1987, we construct a transition matrix across earnings classes for workers with employment during the initial year. From these matrices, the probability of a worker moving to a higher or lower earnings class can be calculated. These changes include both price and quantity changes.

Two main points are readily apparent. First, real earnings are not fixed over time for individuals. Second, the probability of an individual moving downwards in the earnings distribution increases during a recession. Between 1974 and 1975, 34 percent of all workers suffered a decline in real earnings. Similarly, as the recession began between 1981 and 1982, 34 percent of all workers suffered a decline in real earnings. As the recovery got underway between 1983 and 1984, the probability of suffering a real earnings decline fell to 24 percent and remained close to that level as the expansion continued through 1987.

The stable earnings distribution over time hides substantial mobility across earnings classes, mobility that shows procyclical behavior.

V. Conclusion

The disappearance of the middle class does not necessarily follow from industrial restructuring. Pennsylvania represents a dramatic case of industrial restructuring. Between 1974 and 1987, nearly one-third of all manufacturing jobs were eliminated. Despite this, the earnings distribution remained largely unchanged because real wages increased in nearly all sectors. Neither new employees, new enterprises, nor new jobs at growing enterprises show great evidence of an adverse shift in the earnings distribution.

While the industrial restructuring of the early 1980s was large, it was small in relation to the gross employment flows typically observed. Enterprise closings or openings play only a minor role. The economy is quite fluid, with substantial job creation and destruction at ongoing enterprises even during "normal" times.

REFERENCES

Barsky, Robert and Solon, Gary, "Real Wages over the Business Cycle," NBER Working Paper No. 2888, March 1989.

Blackburn, McKinley L., Bloom, David E. and Freeman, Richard B., "Why Has the Economic Position of Less-Skilled Male Workers Deteriorated in the United States?," Brookings Papers in Economics, March 1989.

Bluestone, Barry, "The Changing Nature of Employment and Earnings in the U.S. Economy, 1963–1987," April 1989.

______ **and Harrison, Bennett,** "The Growth of Low-Wage Employment: 1963–86," *American Economic Review Proceedings*, May 1988, *78*, 124–28.

Bound, John and Johnson, George, "Changes in the Structure of Wages during the 1980s: An Evaluation of Alternative Explanations," NBER Working Paper No. 2983, May 1989.

Davis, S. J. and Haltiwanger, J., "Gross Job Creation, Gross Job Destruction and Employment Reallocation," unpublished paper, 1989.

Dunne, Timothy, Roberts, Mark and Samuelson, Larry, "Plant Turnover and Gross Employment Flows in the U.S. Manufacturing Sector," *Journal of Labor Economics*, January 1989, *7*, 48–71.

Horrigan, Michael and Haugen, Stephen, "The Declining Middle-Class Thesis: A Sensitivity Analysis," *Monthly Labor Review*, May 1988, *111*, 3–13.

Jacobson, Louis, "Structural Change in the Pennsylvania Economy, 1979–1985," unpublished paper, Upjohn Institute, 1989.

Juhn, Chinhui, Murphy, Kevin M. and Pierce, Brooks, "Wage Inequality and the Rise in Returns to Skill," unpublished, University of Chicago, October 1989.

Katz, Lawrence and Summers, Lawrence, "Can Inter-Industry Wage Differentials Justify Strategic Wage Policy?," NBER Working Paper No. 2739, October 1988.

Kosters, Marvin H. and Ross, Murray N., "The Distribution of Earnings and Employment Opportunities: A Reexamination of the Evidence," in *Studies in Economic Policy*, Washington: American Enterprise Institute, 1987.

Leonard, Jonathan S., "On the Size Distribution of Establishment and Employment," NBER Working Paper No. 1951, 1986.

______, "In the Wrong Place at the Wrong Time: The Extent of Frictional and Structural Unemployment," in K. Lang and J. Leonard, eds., *Unemployment and the Structure of Labor Markets*, Oxford: Basil Blackwell, 1987.

______, "Firm and Establishment Growth and Stability," unpublished University of California-Berkeley, 1988.

Mitchell, Daniel J. B., "Wage Pressures and Labor Shortages: The 1960s and 1980s," *Brookings Papers on Economic Activity*, 2:1989, 191–231.

Murphy, Kevin M. and Welch, Finis, "Wage Differentials in the 1980s: The Role of International Trade," Unicon Research Corp., Los Angeles, CA, September 1988.

The Impact of Schooling and Industrial Restructuring on Recent Trends in Wage Inequality in the United States

By BARRY BLUESTONE*

Sufficient evidence now exists to confirm a strong rising trend in wage inequality in the United States since at least the mid-1970s. Still elusive is a definitive explanation for this tendency. In the pursuit of one, researchers have now begun to turn their attention to two factors: 1) on the supply side, an apparent rapid growth in the rates of return to education; and 2) on the demand side, an acceleration in industrial restructuring, in particular, the shift in employment from goods production to services. In this connection, the decline in unionization and the expanded exposure of the manufacturing sector to international competition have received renewed consideration. This paper contends that it is precisely the increasing returns to schooling in both goods and services *in combination with* the shift in employment between the two that is responsible for the recent growth in wage dispersion.

I. The Growth in Wage Dispersion

Some of the earliest evidence demonstrating increases in wage inequality, at least for men, can be found in Peter Henle and Paul Ryscavage (1980) and in Martin Dooley and Peter Gottschalk (1984). The former used the Gini coefficient of annual earnings as their indicator of wage dispersion, while the latter relied on the variance in the natural logarithm of annual earnings ($VAR \ln WAGE$). More recent research by Gary Burtless (1989) shows a nearly monotonic increase in Gini coefficients for men beginning soon after World War II, and continuing all the way through the end of his data series in 1986. Burtless' pattern of Gini coefficients for women over the same period (1947–87) resembles a long wave—rising from the end of the war through the early 1960s, declining from a peak of .468 in 1963 to a minimum of .429 in 1980, and then rising again to .442 by 1986.

Adopting the $VAR \ln WAGE$ as the preferred indicator of wage inequality, in that it is more sensitive to changes in the density of the tails of the distribution (see N. C. Kakwani, 1980), Bennett Harrison et al. (1986) found that when men and women are combined in a single series proxying for the entire jobs distribution, wage inequality appears to have declined significantly from 1963 through 1978 but then reversed, tracing out a U-turn in the pattern of dispersion. They report that by 1987, the $VAR \ln WAGE$ had returned to a level higher than that of twenty years earlier. Separating men and women produces a secularly rising $VAR \ln WAGE$ for men à la Burtless and a shallow U-turn for women (see my 1989 paper). When the analysis is restricted to year-round, full-time workers, in order to mitigate the potentially confounding effect of the variance in annual hours worked on earnings, a U-turn with a minimum point in 1975 is found and U-turns exist for men and women analyzed separately.

In analyzing these inequality trends, both Burtless and my earlier paper found virtually no evidence for a business cycle explanation, nor for one based on demographic shifts, despite the increase in female labor force participation and despite changes in the relative size of individual labor force age cohorts. What the more sophisticated econometric evidence reveals is that measured wage dispersion has increased through both macroeconomic contractions and expansions, and it has grown for both men and women, and for both young and old. Something more "structural" must be going on.

---

*University of Massachusetts, Boston, MA 02125.

II. The Variance in Earnings vs. the Variance in Education

The approach taken here is to compare changes in the $VAR \ln WAGE$ of year-round full-time workers with changes in the variance in education and changes in the share of employment in goods production over three periods of time that roughly correspond to the 1963–87 U-turn in wage inequality.

A marginal productivity/human capital analysis would suggest that the trend in wage inequality should be mirrored in similar trends in education. The decline in the variance in earnings during the early period of the U-turn should match a decline in the variance of schooling while the succeeding increase in wage dispersion should reflect an *increase* in the variance in years of school completed.

The evidence does not fit the theory. Table 1 displays the variance in the logarithm of earnings, the coefficient of variation ($\sigma/\bar{X}$) in years of school completed, and the ratio of goods employment to total employment for all year-round full-time workers age 16 and over at four different points in the U-turn: 1963, 1973, 1979, and 1987. From 1963 through 1973, both the variance in education and the variance in earnings dropped substantially consistent with marginal productivity and human capital theory.

However, during the two subsequent periods, wage dispersion rises while inequality in formal schooling continues to decline, albeit at an increasingly slower rate. Overall, between 1973 and 1987, the log of the variance of annual earnings grew by 36 percent in spite of a 15 percent decline in school inequality. Jointly, the two statistics imply a substantial increase in the rate of return to formal schooling.

This finding is consistent with the "job competition" model developed by Lester Thurow (1975) in which an individual's education helps to determine that individual's *relative*, but not absolute, position in the labor market (and wage) queue. According to the theory, if the distribution of wages across existing jobs is determined at least partly exogenously with respect to the supply characteristics of labor, a shift toward a more equal distribution of schooling need not have any direct effect on the distribution of earnings.

TABLE 1—WAGE DISPARITY, EDUCATION INEQUALITY, AND INDUSTRY EMPLOYMENT
(Year-Round Full-Time Workers)

	1963	1973	1979	1987
$VAR \ln WAGE$	.559	.461	.505	.627
$\sigma/\bar{X}$ School	.274	.240	.215	.204
Goods/Total Employment	.394	.370	.336	.288

Source: Annual Demographic Files, March *Current Population Surveys.*

TABLE 2—WAGE DISPARITY, EDUCATION INEQUALITY AND INDUSTRY EMPLOYMENT BY GENDER AND INDUSTRY
(Year-Round Full-Time Workers)

	Percent Change		
	1963–73	1973–79	1979–87
All Workers			
$VAR \ln WAGE$	−17.5	9.5	24.2
$\sigma/\bar{X}$ School	−12.4	−10.4	−5.1
Goods/Total	−6.1	−9.2	−14.3
Men			
$VAR \ln WAGE$	−22.8	12.3	24.5
$\sigma/\bar{X}$ School	−12.0	−10.5	−4.6
Goods/Total	−1.6	−7.5	−10.6
Women			
$VAR \ln WAGE$	−16.3	15.9	42.1
$\sigma/\bar{X}$ School	−10.6	−9.1	−3.1
Goods/Total	−13.4	−7.0	−16.8
Goods Production			
$VAR \ln WAGE$	22.5	−1.4	22.3
$\sigma/\bar{X}$ School	−9.4	−11.1	0.1
Services			
$VAR \ln WAGE$	−12.0	8.4	34.2
$\sigma/\bar{X}$ School	−14.9	−9.7	−6.0

Source: See Table 1.

Table 2 demonstrates that the pattern of increasing wage disparity in the face of a declining variance in education holds true for both men and women. Here, the percentage change in the dispersion measures are reported. Again, during the first phase of the U-turn, earnings inequality tracks educational inequality. During the second and third phases, however, the education/earnings nexus is broken. Nonetheless, there is still some relationship between schooling and wages given that the growth in earnings inequality after 1973 is much sharper for women than men, consistent with smaller declines in educational inequality. The $VAR \ln WAGE$ increased more for women than men in both the 1973–79 and 1979–87 periods, while the coefficient of variation for schooling declined faster for men in both periods, presumably moderating the increase in wage inequality for them.[1] Hence, while the historical direction of the earnings inequality *differential* between men and women is consistent with the standard marginal productivity/human capital argument, the *levels* of wage inequality are not.

A major clue to the wage dispersion puzzle is found when the data are analyzed separately for the goods producing sector and the services sector.[2] In both sectors, wage inequality increases after 1979. But the growth in earnings inequality in services is significantly greater than that in the goods producing sector, in spite of the fact that school inequality actually falls faster among workers in services. The difference is particularly striking when the entire post-1973 experience in the two sectors is compared (not shown in table.) The increase in $VAR \ln WAGE$ in services is more than double the increase in goods production (45.4 vs. 20.6 percent), while education was becoming proportionately more equally distributed in services.[3]

Combining this result with the fact that goods production employment has declined from 39 percent of the total in 1963 (and 37 percent in 1973) to only 29 percent in 1987 provides a final clue. As labor was displaced from the goods production sector and swelled the ranks in the service industries, workers moved from what is essentially an industrial sector with higher mean earnings and a lower wage variance to a sector with the opposite characteristics: a lower mean wage and a higher variance. In brief, industrial restructuring not only has contributed to lower annual earnings but to higher wage inequality.

The findings in Table 2 for men and women provide corroborative evidence for the above conclusion. Given popular perceptions about "male-dominated" construction trades and manufacturing industries, one might conjecture that the earlier finding that women have experienced a higher increase in wage inequality would be contrary to the industrial restructuring argument. The common impression is that men have been the primary victims of "deindustrialization," insinuating that they should have borne the brunt of the wage dispersion impact of the shifts from goods to services. As Table 2 indicates, the popular perception is incorrect. The percentage shift from goods production employment to services employment is actually greater for women than for men, particularly after 1979. Between 1973 and 1987, the ratio of goods employment to total employment declined by 17 percent for men but by nearly 23 percent for women. This seeming paradox is explained not only by the extraordinary growth in service employment for women, but by the fact that the largest declines in manufacturing employment came in sectors that have been traditionally female dominated: apparel, textiles, and shoes. Moreover, the common practice of ordering layoffs according to seniority in manufacturing has likely generated the same gender difference in the incidence of displacement from blue-collar durable goods production.

III. Some Thoughts on Explaining the Trends

The interaction between a rising rate of return to education in both sectors plus the shift in employment from goods to services is suggestive of the mechanism by which overall wage variance is increasing, but by itself begs the question as to why either of these trends has materialized. Here I can briefly conjecture about both.

---

[1] Not shown in the table are the actual variances which increased between 1973 and 1987 from .414 to .579 for men and from .365 to .601 for women.

[2] The "goods" sector here includes construction and durable and nondurable manufacturing. The "services" sector includes transportation, communications, and public utilities; wholesale and retail trade; finance, insurance, and real estate; business services; personal services; entertainment and recreation services; and professional services. Agriculture, mining, and government service are excluded from this analysis.

[3] This is shown in the actual *levels* of school inequality in the two sectors. In 1963, the coefficient of variation in schooling among workers in the goods sector and in the service sector was nearly identical: .267 in the former vs. .270 in the latter. By 1987, education was substantially more equally distributed in services: .195 vs. .216 in goods production.

The shift toward services and out of goods production is perhaps the easier to explain. It presumably reflects the impact of: (a) import competition in the manufacturing sector; and (b) faster increases in productivity growth in goods production.[4] Both of these offer reasons why in the period since 1973, the decline in goods production employment has been so precipitous. It also suggests one reason why Kevin Murphy and Finis Welch (1988) find that among all of the variables they test, trade deficits in durable manufactured goods is the single best explanation for the growing earnings gap since 1978 between white men with a high school diploma or less and those with the college degree or more.

As for the growth in inequality *within* the goods sector, one can point to (a) the decline in unionization within manufacturing, (b) the shift of part of the manufacturing sector toward lower wage regions of the country, and (c) internal restructuring toward wage systems based on wage concession bargaining, the imposition of two-tier wage structures, the substitution of contingent labor for full-time regular workers, and the outsourcing of jobs to smaller contractors that often have lower pay scales (Harrison and myself, 1988). Declining unionization in particular plays a significant role in explaining the earnings gap between high school dropouts and college graduates in the work of McKinley Blackburn et al. (1989).

The growth in inequality within the service sector is more difficult to explain, particularly in light of the disproportionately larger decline in education inequality relative to the goods sector. The extremely low union density in services might explain why the *level* of wage dispersion is higher than in goods production, but this alone would not account for the *change* in wage dispersion over time. The one possibility that needs further investigation is that with the rapid expansion in services, union density may actually have fallen faster than that in goods production.

IV. Conclusion

In summary, the research begun only a few years ago into the growing dispersion in wages in the United States has already begun to bear fruit. Few any longer doubt the nature of the trends in inequality, and there is increasing consensus on at least some of the root causes.

We now know that inequality has increased for both men and women since at least the early 1970s, and that it has increased in both the goods production and service sectors. There is broad consensus that little of the trend can be explained simply by demographic factors or by the business cycle. Similarly, changes in the distribution of education can at best explain only a fraction of the overall trend.

Greater agreement is now found in theories that focus on the demand side of the market—on industrial restructuring, on the decline in unionization, and on the impact of international trade particularly on the manufacturing sector. The major unanswered question concerns what the future holds. Will the trends toward greater wage dispersion continue or begin to subside in the 1990s? If the current research is to be believed, the answer to this question will be determined by such factors as America's competitive position in the world economy and on the fate of unionization.

REFERENCES

Baily, Martin Neil and Blair, Margaret M., "Productivity and American Management," in Robert E. Litan, et al., eds., *American Living Standards: Threats and Challenges*, Washington: Brookings Institution, 1988.

---

[4] Martin Neil Baily and Margaret Blair (1988) report that from 1979 to 1986, growth of GDP per hour in the manufacturing sector grew by 3.10 percent per year while the growth of GDP per hour for the total economy was only 1.09 percent.

Blackburn, McKinley, Bloom, David and Freeman, Richard B., "Why has the Economic Position of Less-Skilled Male Workers Deteriorated in the United States?," Brookings Discussion Papers in Economics, March 1989.

Bluestone, Barry, "The Changing Nature of Employment and Earnings in the U.S. Economy: 1963–1987," Conference on Job Creation in America, University of North Carolina-Chapel Hill, April 1989.

Burtless, Gary, "Earnings Inequality over the Business Cycle," Brookings Discussion Papers in Economics, February 1989.

Dooley, Martin and Gottschalk, Peter, "Earnings Inequality among Males in the United States: Trends and the Effect of Labor Force Growth," *Journal of Political Economy*, February 1984, *92*, 59–89.

Harrison, Bennett and Bluestone, Barry, *The Great U-Turn: Corporate Restructuring and the Polarizing of America*, New York: Basic Books, 1988.

______, ______, and Tilly, Chris, "Wage Inequality Takes a Great U-Turn," *Challenge*, March/April 1986, *29*, 26–32.

Henle, Peter and Ryscavage, Paul, "The Distribution of Earned Income Among Men and Women, 1958–1977," *Monthly Labor Review*, April 1980, *103*, 3–11.

Kakwani, N. C., *Income Inequality and Poverty: Methods of Estimation*, Oxford: Oxford University Press, 1980.

Murphy, Kevin and Welch, Finis, "The Structure of Wages," University of Chicago Working Paper, August 1988.

Thurow, Lester, *Generating Inequality*, New York: Basic Books, 1975.

Schooling, Work Experience, and Wage Trends

By MARVIN H. KOSTERS*

When members of the baby-boom generation began entering the work force in substantial numbers during the 1970s, the resulting crowding reduced their earnings in comparison with earlier cohorts of the same age, and wage premiums for schooling also declined. These developments raised questions about how well members of the baby-boom generation might fare during their working lives compared with smaller neighboring cohorts (Finis Welch, 1979; Mark Berger, 1985). When wage premiums for schooling increased sharply during the 1980s, however, it became apparent that workers' levels of schooling were more important determinants of their lifetime earnings than the size of their birth cohort. The small decline in schooling wage premiums during the 1970s now appears to have been only a temporary interruption of a trend toward higher economic rewards for schooling.

Since schooling and work experience account for about half of the variation in earnings, wage trends of the work force as a whole have been influenced by changes in its composition since the late 1960s. In addition to these compositional changes, relative wages have also changed dramatically as the number of entrants into the labor force with high schooling levels surged during the 1970s and receded during the 1980s. The purpose of this paper is to describe the relationship between changes in the schooling and work experience composition of the labor force and changes in relative wages during the past fifteen years.

I. Average Wage Levels

To set the stage, it is useful to begin with a note on the data. Hourly wage rates by schooling and work experience categories are derived from *Current Population Survey* information. Five schooling classes and five-year potential work experience categories were defined for males and females separately. Hourly earnings reported by workers paid by the hour were used as their wage measure, and wages of others were computed from their usual weekly hours and earnings. For observations affected by weekly earnings top codes, estimates were constructed based on the Pareto distribution, with parameters estimated separately for each year by schooling class and sex. Wage data were adjusted for inflation by using CPI-U-X1.

The slower growth in the average worker's real pay from the 1950s and 1960s to the 1970s and 1980s is widely recognized. Total real compensation per hour of work increased by less than 1 percent per year during the fifteen years between 1973 and 1988, for example, compared with more than 2 percent during the preceding twenty years. Supplements to hourly wages (mainly larger contributions for pension and health plans but also cash bonuses) have grown as a share of total compensation, however, and hourly earnings have accordingly increased less rapidly than total pay. Some measures, such as hourly earnings of production and nonsupervisory workers (from establishment reports), show a decline in real wages compared with their 1973 peak. Other measures of real hourly earnings for the work force as a whole show very little net change from 1973 to 1988. All these measures depend on the price index used to account for inflation and on information availability. Generally, only data on the wage and salary component of total compensation are available for comparisons over time for particular demographic groups.

After declining in the 1970s, real average hourly earnings for both men and women rose by more than 10 percent during the 1980s. The small net increase in real wages from 1973 to 1988 of about 6.5 percent for both sexes was almost entirely attributable to rising schooling levels. Wage indexes calculated using fixed 1973 schooling and work experience weights indicate that, by 1988, wages would have been about 10 percent below actual levels for both males and females in the absence of increased schooling. Changes in work experience composition produced very small negative effects in the 1970s, followed by small positive contributions in the 1980s.

---

*Director, Economic Policy Studies, American Enterprise Institute, Washington, D. C. 20036.

The increase in the proportion of the work force accounted for by women resulted in an increase in average real wages for the work force as a whole smaller than those for males and females separately, about 4 percent. Although arithmetic means show essentially no change in relative hourly earnings for men and women, geometric means show women's hourly wages rising relative to men's, and earnings of year-round full-time workers also show somewhat larger gains for women. The main effects of changing work force composition on average wage levels came from increases in the schooling levels of workers and in the share of the work force accounted for by women.

II. Schooling and Wages

During most of the 1970s, the wage premium for additional schooling declined modestly. This decline in the economic returns to schooling, which followed an increase of comparable size during the late 1960s, was concentrated among younger workers with little work experience. This small decline during the 1970s was decisively reversed during the 1980s, however, with schooling wage premiums reaching levels far above those that prevailed earlier.

Several characteristics of the remarkable rise in the economic returns to schooling during the 1980s are worth noting. First, wage premiums rose across the entire spectrum of schooling levels for both men and women. The percentage wage gap between schooling levels widened between all schooling classes. Second, real wage levels for workers with thirteen to fifteen years of schooling remained roughly unchanged throughout the 1970s and 1980s. Thus, for workers with twelve years of schooling or less, real wages declined in the 1980s, while for workers with sixteen years of schooling or more, real wages rose. Finally, economic returns to schooling increased to extraordinary levels during the 1980s. By 1988, wage premiums for college were substantially higher than in the early 1970s.

TABLE 1—COLLEGE/HIGH SCHOOL WAGE PREMIUMS: SELECTED YEARS, 1973–88

	Arithmetic Means		Geometric Means	
	Men	Women	Men	Women
All Ages				
1973	43	48	37	45
1978	35	36	31	35
1980	39	34	32	32
1983	48	39	40	37
1988	78	51	61	49
Ten Years Work Experience				
1973	49	49	44	51
1978	36	38	34	41
1980	31	37	35	38
1983	44	46	45	47
1988	86	59	69	61

Notes: Shown in percent. The source of the data is *CPS* data files using the May *Survey* for 1973–78 and April-May-June *Surveys* for 1979–88. Arithmetic means for 10 years of work experience are averages for workers with 6 to 15 years of potential work experience and geometric means are antilogs of mean log wages predicted by separate regressions by sex, year, and schooling class, evaluated at 10 years of potential work experience.

Estimates of the percentage wage premium for college are reported in Table 1. Both the initial decline in schooling wage premiums and their subsequent sharp rise during the 1980s stand out clearly. The change in the 1980s can be summarized reasonably well with the generalization that college wage premiums rose by about 50 percent for women and doubled for men. The one exception to this generalization is the larger increase for young men shown by estimates based on arithmetic means. The fact that the wage premium for young men based on geometric means approximately doubled suggests an increase in density in the upper tail of their wage distribution during the 1980s.

College wage premiums for young workers with ten years of potential work experience show the combined effects of changes in schooling wage premiums and in experience profiles. The high premiums for young workers shown by all of the estimates for 1988 reflect primarily the steepening of experience profiles for high school-level workers. Relative wages of young workers with no schooling beyond high school declined substantially during the 1980s while experience profiles for workers with college-level schooling remained quite stable.

Young workers with relatively low schooling levels were affected by both the decline in relative wages associated with the rise in schooling wage premiums and the steepening of their experience profile. The combined effects of these two adverse shifts in relative wages was particularly pronounced for males. Real wages declined by 12 percent from 1980 to 1988 for males with high school and ten years of work experience. Male college graduates, however, experienced a 20 percent increase in real wages. The contrast is even sharper for workers who did not complete high school compared with those with schooling beyond the college level.

Research has been conducted to identify the main sources of the rise in schooling wage premiums in the 1980s. Changes in demographic composition and in aspects of labor quality that are not captured by measures of schooling and potential work experience, such as immigration, have been analyzed as factors influencing supply. Changes in demand patterns, such as shifts in the composition of product demand and technological change have been studied. Changes in market conditions, such as increased competition from abroad, shifts from internally produced to purchased components, and changes in factors influencing wage setting, such as the decline in unionization and in the real minimum wage, have also been examined (John Bound and George Johnson, 1989; Lawrence Katz and Ana Revenga, 1989; and McKinley Blackburn et al., 1989).

This research has contributed to a better understanding of the possible sources of changes in relative wages. Since many of these factors may be closely interrelated, however, their potential effects remain inconclusive. Although I recognize that the underlying causes of the rise in schooling wage premiums during the 1980s have not been established, I argue on the basis of internal evidence on changes in relative wages and changes in rates of growth in the supply of schooling that the observed trends are at least qualitatively consistent with strong growth in relative demand for schooling and skills.

III. Schooling and Relative Supply

Even though schooling levels for the work force as a whole continued to rise during the 1980s, the economic return to schooling rose sharply. Unless schooling is extremely inadequate as a measure of economically important attributes of labor supply, rising schooling wage premiums at a time when schooling levels in the work force are being upgraded cannot be consistent with stable, proportionate growth in labor demand. This conclusion is persuasively demonstrated in a thorough analysis by Kevin Murphy and Welch (1988).

If demand for workers with more skills and schooling has increased over the years relative to that for workers with less schooling, changes in the rate of growth in supply can help to explain changes in schooling wage premiums. Schooling levels were being upgraded much more rapidly during the 1970s, when schooling wage premiums declined modestly, than in either the 1960s or the 1980s. The slower upgrading of schooling levels during the 1980s was particularly evident for men, who also experienced the largest increase in schooling wage premiums.

Consider first how entry of large numbers of young workers changed the age composition of the work force. During the 1960s, the proportion of 25 to 34-year-olds rose by .04 and .06 percentage points per year, respectively, for men and women. During the 1970s, however, the labor force shares for these young workers increased by .55 and .88 percentage points per year. The increases slowed during the 1980s to .15 and .13 per year. Schooling levels of these young workers, which were much higher than those of retiring workers, are reflected by changes in the schooling composition of the work force.

During the 1970s, the share of the work force with high school or less was declining by about 1.25 percentage points per year for men and 1.27 for women. The share with college-level schooling was increasing by .78 percentage points per year for men and .68 for women. These changes were more than twice as large as in the 1960s. During the 1980s, the share with high school or less was declining by only .46 percentage points per year for men and by .90 for women, and shares with college-level schooling were increasing by less than .5 percentage points per year for both men and women. The rate at which schooling levels were being upgraded accelerated sharply during the 1970s, and then slowed substantially during the 1980s.

Three important changes from conditions that prevailed in the 1970s contributed to the decline in the rate at which schooling levels were upgraded in the 1980s. First, although new entrants in the 1980s still had considerably more schooling than retiring workers, the rate of growth of the work force began to taper off as smaller increments of young workers were added to the work force. Second, the schooling composition of new entrants began to stabilize, and for males the proportion with high school or less even rose slightly. Third, demands for replacement of workers with higher levels of schooling increased because the schooling levels of workers approaching retirement in the 1980s were appreciably higher than in the 1970s.

In addition to quantitative differences in supply as measured by levels of schooling, changes in the quality of schooling could also produce changes in schooling wage premiums. If the growing public concern during the 1980s about the quality of elementary and high school education reflects actual deterioration, for example, the rising college wage premium might reflect it. A selectivity bias might be present as a growing fraction of young people complete at least some college training. In the absence of that, however, deterioration in the relative quality of elementary and secondary schooling would be reflected by rising college wage premiums for workers who completed their schooling more recently.

Examination of trends in the college wage premium for different birth cohorts during the 1980s suggests that they were not caused primarily by changes in relative quality of schooling. College wage premiums rose sharply across the age spectrum; there is no systematic evidence that workers who completed their schooling in the 1950s or 1960s were affected any more or less than those who completed their schooling during the 1970s. If schooling quality has changed over the years, it is apparently not the main source of the rise in schooling wage premiums during the 1980s. Contemporaneous labor market forces were instead apparently the dominant source of the change.

IV. Labor Market and Policy Considerations

Trends in the economic return to schooling during the 1970s and 1980s are consistent with more rapid growth in demand for workers with high levels of schooling than for workers with less schooling. The extraordinarily large influx of young workers with higher levels of schooling during the 1970s was accompanied by only a small decline in the economic return to schooling. The slowdown in the rate at which schooling levels were upgraded during the 1980s, however, produced a pronounced rise in wage premiums. These trends suggest that the U.S. economy has a strong capacity for making productive use of workers with large investments in schooling.

Increased schooling wage premiums represent short-term adjustments to skill demands. High economic returns to schooling, if expected to persist, will bring about adjustment in relative supplies as young people are stimulated to acquire more schooling. Trends in college enrollment rates, for example, have been positively related to schooling wage premiums. Job-related training for workers with less schooling is another way of augmenting skill supplies. Their lower relative wages make hiring workers with less schooling more attractive to employers, particularly if prospects for retaining such workers are favorable enough so that employers can expect to recoup the cost of investments in training that may be necessary.

Policies that would foster labor market adjustments to strong skill demands could also be considered. To strengthen employers' incentives to hire and train workers with low schooling levels, for example, management tools such as gradual vesting of employers' contributions to pension plans could be encouraged. To encourage higher college enrollment, availability of loans could be expanded. To improve incentives for achievement while in school, employment testing could be restructured to place more emphasis on general instead of specific job-related knowledge. To make more effective use of skills of older workers, the high implicit taxes that now discourage them from continuing to work could be phased out.

Since increases in average real wage levels were quite small while major shifts in relative wages were taking place, real wages of some workers declined. Incomes of young workers with low schooling levels were most severely affected. Extension of the current 14 percent earned income tax credit to workers without dependents would be a simple and straightforward way of raising after-tax incomes for many of these workers. Incomes of workers with dependents could be augmented by raising personal exemption levels for dependents or by a dependent tax credit. Policies structured along these lines to support incomes of workers with low earnings would be more constructive than efforts to protect their wages against declines in an increasingly competitive world economy.

REFERENCES

Berger, Mark C., "The Effect of Cohort Size on Earnings Growth: A Reexamination of the Evidence," *Journal of Political Economy*, June 1985, *93*, 561–573.

Blackburn, McKinley, Bloom, David and Freeman, Richard B., "Why Has the Economic Position of Less-Skilled Male Workers Deteriorated in the United States?," Brookings Discussion Paper on Economics, March 1989.

Bound, John and Johnson, George, "Wages in the United States During the 1980s and Beyond," American Enterprise Institute Conference Paper, November 3, 1989.

Katz, Lawrence F. and Revenga, Ana L., "Changes in the Structure of Wages: The U. S. Versus Japan," NBER Working Paper No. 3021, July 1989.

Murphy, Kevin and Welch, Finis, "The Structure of Wages," Unicon Research Corporation, Los Angeles, April 1988.

Welch, Finis, "Effects of Cohort Size on Earnings: The Baby Boom Babies' Financial Bust," *Journal of Political Economy*, October 1979, Part 2, *87*, 565–97.

SELECTIVITY BIAS: NEW DEVELOPMENTS[†]

Varieties of Selection Bias

By JAMES HECKMAN*

This paper considers a prototypical problem in the econometrics of selection bias: estimating the impact of unionism on wage differentials. The statistical structure of this problem is the same as that of many other self-selection problems in economics. The union wage differential question is of interest in its own right, and has been the subject of numerous papers evaluated in the landmark study of H. Gregg Lewis (1986).

I consider the following questions. 1) What are the parameters of economic interest? 2) Under what conditions can they be identified? 3) How can alternative nonparametric procedures set forth in the literature aid in securing the parameters of interest? 4) What is the status of the evidence on appropriate methods for estimating union-nonunion wage differentials? The answer to the third question is particularly relevant in view of Lewis' pessimistic and influential summary of existing studies that attempt to correct for self-selection bias. He complains about the wide variation in the estimates produced in different studies. Such variability may be due to the imposition of false distributional assumptions—an issue that motivates the recent literature on nonparametric estimation.

I. What Are the Parameters of Economic Interest?

Which parameters are we seeking to estimate without self-selection bias? Suppressing individual subscripts, a switching regressions model of unionism writes union wages Y_1 as $Y_1 = X_1\beta_1 + U_1$, $E(U_1) = 0$; nonunion wages Y_0 as $Y_0 = X_0\beta_0 + U_0$, $E(U_0) = 0$; and the sectoral choice equation as $I = Z\gamma + V$, $E(V) = 0$; where $I \geq 0$ implies that union sector one is chosen ($D = 1$). Otherwise $D = 0$.

The observed wage Y is

$$(1)\quad Y = 6Y_1D + Y_0(1 - D)$$
$$= (X_1\beta_1)D + (X_0\beta_0)(1 - D)$$
$$+ DU_1 + (1 - D)U_0.$$

The case $I = Y_1 - Y_0$ is the Roy model, put in linear regression form. Lung Fei Lee (1978) considers more general self-selection rules. Following much of the literature on this problem, assume that (U_1, U_0, V) are independent of the regressors (X_1, X_0, Z). In the original models, (U_1, U_0, V) were assumed to be joint normal.

The most commonly used specification of this model assumes that $\beta_1 = \beta_0$ except for the intercepts α_1, α_0, and $X_1 = X_0$. Letting β be the common slope coefficient vector and X the vector of regressors excluding the intercept, we obtain

$$(2)\quad Y = \alpha_0 + D(\alpha_1 - \alpha_0)$$
$$+ X\beta + (1 - D)U_0 + DU_1.$$

The empirical literature on the simultaneity problem in estimating the impact of union status focuses on the stochastic dependence between D and the disturbance $\{U_0 + D(U_1 - U_0)\}$. But the parameters of interest are often not well defined, and are also defined differently in various studies—sometimes inconsistently in the same study.

[†] *Discussants*: Paul Rudd, University of California-Berkeley; Whitney Newey, Princeton University; Roger Klein, Bell Communications Research.

*Department of Economics, Yale University, 37 Hillhouse, New Haven, CT 06520. This research was supported by NSF grant no. 87-11845. Steve Cameron, Kyung Soo Choi, Bo Honoré, Whitney Newey, and James Walker gave helpful comments.

Most economists have been content with estimating an average effect. But which average? Two have been suggested in the literature. Sometimes they have been confused. Under certain conditions, they are the same. When they are not, they differ in the assumptions that must be invoked to identify them in the presence of various selection rules.

The first average is the experimental treatment average: what is the effect of randomly (i.e., as in an ideal laboratory experiment) moving a worker from the nonunion sector to the union sector? In terms of equation (2), this parameter is

$$(\alpha_1-\alpha_0). \tag{3}$$

In the more general model, it is often defined as

$$\tilde{X}_1\beta_1-\tilde{X}_0\beta_0, \tag{4}$$

where $\tilde{X}_1$, $\tilde{X}_0$ are the endowments of the skills in each sector assumed available to the "average" worker in the unionized sector $(\tilde{X}_1=E(X_1|D=1)$, $\tilde{X}_0=E(X_0|D=1))$.

Despite the focus of most of the attention on these parameters in the recent empirical literature, they do not always answer economically interesting questions. It is also interesting to know what is the effect of unionism on the unionized. In the context of equation (2), the parameter is

$$\begin{aligned} E(Y_1-Y_0|D=1,x,z) \\ =(\alpha_1-\alpha_0)+E(U_1-U_0|D=1,z). \end{aligned} \tag{5}$$

This is the gain from moving a unionized person with attributes $X=x$ and $Z=z$ from the nonunionized to the unionized sector. This is what Lewis calls the wage gap (p. 11), although he does not consistently define it this way throughout his survey. If $E(U_1-U_0|D=1,z)=0$, the two parameters are equivalent and instrumental variables estimators of $(\alpha_1-\alpha_0)$ are consistent. (See my paper with Richard Robb, 1985; Gregg Duncan and Duane Leigh, 1985; and Christopher Robinson, 1989).

The distinction between (3) or (4) and (5) is crucial in assessing alternative selection bias estimators, and in evaluating union wage impacts. Yet these parameters are often confused. The wide variability in estimates that causes Lewis to dismiss econometric methods for solving self-selection problems arises in part because parameters (3) and (5) are confused in his comparison.

II. What Can Be Identified?

Recent advances in econometrics have focused attention on nonparametric and semiparametric estimation of economic models. Identification theorems are a necessary first step in this process, and provide useful discipline in separating out parameters of interest that can only be identified by invoking arbitrary functional form assumptions from those that are nonparametrically identified. In this section, I prove that, under conventional assumptions about sectoral wage and selection equations, a version of the Roy-Lee model of unionism is nonparametrically identified. Actually a more general model can be identified. To prove this, I adopt a slightly more general notation. Separate out the variables in common with X_1, X_0, and Z and call these X_c. The variables left over are relabeled X_1, X_0 and Z. Let

$$Y_1=g_1(X_1,X_c)+U_1,$$

$$Y_0=g_0(X_0,X_c)+U_0,$$

$$I=(Z,X_c)\gamma+V$$

where $I\geq 0$ implies $D=1$, and $D=0$ otherwise.

The following theorem can be proved. (It is based on a modification of a theorem in my paper with Bo Honoré, 1990.)

THEOREM: *Let* (U_1,U_0,V) *be median (or mean) zero i.i.d. random variables with density* $f(U_1,U_0,V)$ *distributed independently of* $X=(X_1,X_0,Z,X_c)$. $\mathrm{Var}(V)=1$. *If*

(A) (U_1,U_0,V) *are continuously distributed with distribution function F and support equal to* R^3;

(B) *Support* $((Z,x_c)\gamma)=R^1$ *for all* x_c *in the support of* X_c. *There is no proper linear subspace of the space generating* (Z,X_c) *hav-*

ing probability 1 *with respect to the distribution of* (Z, X_c);

(C) *The marginal distributions of* (U_1, U_0, V) *have medians* (*means*) *equal to* 0. *Then* $F(u_1, v)$, $F(u_0, v)$, $g_1(x_1, x_c)$, $g_0(x_0, x_c)$ *and* γ *are identified from the distributions* $F(y_1|D=1, x_1, x_c, z)$, $F(y_0|D=0, x_0, x_c, z)$ *and the distribution of D*, $Pr(D=1|z, x_c)$.

PROOF:

Using $Pr(D=1|z, x_c)$ it is possible to mimic Charles Manski's (1988) proof of the identifiability of γ and $F(v)$ in the binary choice model. Using the distribution of sampled Y_1 and D, $Pr(Y_1 \le y_1|D=1,\ X_1=x_1,\ X_c=x_c,\ Z=z) = Pr(g_1(x_1, x_c)+U_1 \le y_1,\ V \ge -(z, x_c)\gamma)/Pr(V \ge -(z, x_c)\gamma)$. Fix $X_c = x_c$. Define an isoprobability set for Z as $Z_p = \{z\colon Pr(D=1|z, x_c)=p\}$. For each element in this set, define another set

$$\left(\bar{\bar{Y}}_1, \bar{\bar{X}}_1\right)_{q,p} = \{(y_1, x_1)\colon Pr(Y_1 \le y_1|D=1,$$

$$X_1=x_1,\ X_c=x_c,\ Z=z) = q$$

$$= Pr(g_1(x_1, x_c)+U_1 \le y_1,$$

$$V \ge -(z, x_c)\gamma)/Pr(V \ge -(z, x_c)\gamma)\}.$$

Using all x_1 and y_1 pairs within this set, it is possible to recover g_1 up to an additive constant. From knowledge of g_1 up to a constant, it is possible to trace out the distribution of (U_1, V) up to a constant for U_1 for each value of p, using all values of q. Using (C), it is possible to identify the constant. Since x_c was arbitrary this completes the proof for $g_1, \gamma, F(u_1, v)$. By a parallel argument, g_0 and $F(u_0, v)$ can be identified.

To identify β_1, β_0, augmented to be the coefficients of (X_1, X_c) and (X_0, X_c), respectively, it is necessary to assume that there is no proper linear subspace of the space generating (X_1, X_c) and (X_0, X_c) with probability one with respect to the distributions of these random variables. Under the conditions of the theorem, it is thus possible to nonparametrically identify $E(U_1-U_0|D=1, z, x_c)$, and hence it is possible to nonparametrically identify the effect of unionism on the unionized

$$E(g_1-g_0+U_1-U_0|D=1, z, x).$$

The crucial feature of (B) is the presence of a regressor in Z that traces out the probability of selection. The g_1 and g_0 functions can be constants as in the classical Roy model. The argument fails if we permit the distribution of U to depend on X in a general way. However, it is obviously possible to permit the distribution to depend on X_c in a general way. The theorem can be generalized in many ways. Normalizations other than $\mathrm{Var}(V)=1$, such as $\gamma'\gamma=1$, are clearly possible. The support of either U_1 or U_0 need not be R^1. Different variables may be in common in the Y_1, I equations than in the Y_0, I equations.

The intuition underlying this theorem is very simple. From Manski, the parameters of the sectoral choice probability $(Pr(D=1|z, x_c))$ can be identified. For each value of the sectoral choice probability (p), stratify the data on y_1 and (x_1, z, x_c). Fixing (z, x_c) so that $Pr(D_1=1|z, x_c)=p$, we may trace out a set of (x_1, y_1) values such that $Pr(Y_1 \le y_1,\ V \ge -(z, x_c)\gamma)=q$. For each (x_1, x_c) in the set, we find the associated y_1. This identifies g_1 up to a constant c. Transforming the data by $y_1-(g_1+c)$, we can trace out the distribution $F(u_1, v)$. Doing this for all x_c, p, and q, using the mean (or median) zero assumption to identify c, and repeating the exercise for Y_0, we can identify the stated parameters of the theorem.

Note that, under the conditions of the theorem, it is not possible to identify the full joint distribution $F(u_1, u_0, v)$. This is intuitively reasonable since, by hypothesis, we never observe Y_1 and Y_0 for the same person. Placing a restriction on the decision rule such as the hypothesis that sectoral choices are made on the basis of income maximization, I becomes

$$I = g_1(X_1, X_c) - g_0(X_0, X_c) + U_1 - U_0.$$

Honoré and I establish under the conditions of the theorem that the joint distribution $F(u_1, u_0)$ can be nonparametrically identified so that it is possible to identify dependence between Y_1 and Y_0.

Identification results of the sort reported here are very fragile. If the Z regressors are finite valued, identification fails. The continuity of the underlying distribution and the assumption that $(z, x_c)\gamma$ can be varied over the real line play a crucial role in securing identification.

III. The Current State of the Art in Semiparametric Estimation of Selection Models

Identification is only a necessary first step toward estimation. Important progress on the consistent nonparametric estimation of selection models has been made by Stephen Cosslett (1990), Ronald Gallant and Douglas Nychka (1987), Hidehiko Ichimura and Lee (1990), Don Andrews (1989), Whitney Newey (1988), and James Powell (1989). It is interesting to compare what is theoretically possible, and economically interesting to know, with what has been produced in the recent literature. The models differ in the assumptions made and results that can be obtained. All authors except Gallant-Nychka assume that g_1 and g_0 are linear functions: $g_1 = (X_1, X_c)\beta_1$, $g_0 = (X_0, X_c)\beta_0$. Gallant-Nychka make smoothness and continuity assumptions about the distribution of the errors stronger than those used in my identification theorem, and produce consistent nonparametric estimators of the densities $f(u_1, v)$ and $f(u_0, v)$ and the parameters β_1, β_0 and γ. They do not produce a distribution theory for their estimator.

The other papers do not make strong independence assumptions and take as their point of departure the conditional mean-index-function representation of the sample selection problem. Focusing on the equation for Y_1,

$$(6)\quad E(Y_1|D=1, x_1, x_c, z) = (x_1, x_c)\beta_1 + E(U_1|D=1, z, x_c),$$

where $E(U_1|D=1, z, x_c) = \phi((z, x_c)\gamma)$ and where $(z, x_c)\gamma$ is an index function (minus the inverse function of $Pr(D=0|z, x_c)$ when V is independent of $(Z, X_c)\gamma$). The dependence of U_1 on Z and X_c comes strictly through the index. The objective of these papers is to eliminate the contaminating effect of $E(U_1|D=1, z, x_c)$ in forming regression estimates of β_1. Ichimura-Lee use the index property to jointly estimate β_1, γ by kernel methods. They assume access only to truncated samples and hence do not assume access to information on γ or $Pr(D=1|z, x_c)$ in forming their estimator. All of the other authors do. Andrews and Newey approximate the conditional mean of U_1 by series expansion methods assuming censored samples. Cosslett exploits the index function property and approximates $E(U_1|D=1, z, x_c)$, using step functions based on his nonparametric estimator of γ and the distribution of V. Powell uses kernel function methods and assumes censored sampling, and uses the information on γ to form approximate equivalence classes within which $Pr(D=1|z, x_c)$ and hence $E(U_1|D=1, z, x_c)$ are approximately constant. Within the approximate equivalence classes, he differences out the conditional mean. As sample size increases and kernel window widths contract, the approximation becomes arbitrarily good. All of these authors except Cosslett produce a large sample distribution theory for their estimators.

These papers present various identification criteria. The most agnostic assume truncated sampling so there is no information available to estimate $Pr(D=1|z, x_c)$. Ichimura-Lee forego identification of components of β_1 that are associated with variables in the index function. "ϕ" may be linear in $(z, x_c)\gamma$ and so such components are not identified. Models that exploit the index property more fully and assume access to censored samples require that there be one regressor in Z and that $\phi((z, x_c)\gamma)$ does not lie in the space spanned by X_1, X_c. Both groups of models absorb the intercept into the definition of $E(U_1|D=1, z, x_c)$. Thus *none* of these papers except for Gallant-Nychka produces consistent estimators of parameters (3), (4), or (5).

With enough variation in Z and continuity in the densities, it is possible to separate $E(U_1|D=1, z, x_c)$ from $(x_1, x_c)\beta_1$. An index function assumption strengthened with sufficient variation in Z will work. Thus for those Z such that $Pr(D=1|z,x_c)$ becomes arbitrarily close to unity, $E(U_1|D=1,z,x_c)$ becomes arbitrarily close to 0. With sufficient variation in the (X_1, X_c) for this set of Z, β_1 can be identified by least squares. (This is an instance of Gary Chamberlain's 1986 "identification at infinity.") Using the set of (z,x_c) that set $Pr(D=1|z,x_c)$ close to one, it is possible to modify Andrews' and Newey's procedure to consistently estimate the intercepts of equation (6). Rather than estimating all parameters on what may be a small probability set, I recommend a two-step procedure. 1) Exploit the results of Andrews and Newey to estimate the parameters that can be identified from the available distribution of the data. 2) Fixing these parameters, use the values of z and x_c that set $P(d=1|z,\ x_c)$ close to one, and in the limit becomes one, to identify and estimate the remaining parameters. The analogy to this strategy in density estimation is the difference between estimating a mode and a tail of a density. Different points of expansion estimate different parameters. The distribution theory for the second step of this order statistic estimator remains to be developed. With knowledge of $E(U_0|D=0,z,x_c)$ obtained from a parallel analysis of the Y_0 equation, it is possible to use an iterated expectation argument to exploit $E(U_0|z,x_c)=0$ to find

$$E(U_0|D=0,z,x_c)$$
$$=-E(U_0|D=1,z,x_c)\frac{Pr(D=1|z,x_c)}{Pr(D=0|z,x_c)}.$$

From censored samples, it is in principle possible to estimate (3), (4), and (5). It is not necessary to invoke full independence assumptions. Index dependence of the conditional mean suffices. However, it is necessary to exploit more information than is currently used in any paper—except that of Gallant-Nychka. They replace identification at infinity with smoothness and independence conditions that effectively enable them to exploit the benefits of identification at infinity. Full independence allows the analyst to extrapolate out of the sample and to compute the full distribution and not just the mean of union wage impacts.

I am not necessarily advocating the imposition or use of this type of information. However, if it is not imposed, we abandon the pursuit of many economically interesting questions. If it is used, we know the circumstances under which we can address them. This is a fundamental advance in our knowledge about selection bias models.

IV. Simpler Methods May be Robust After All

As previously noted, if $E(U_1-U_0|D=1,z,x_c)=0$, instrumental variables estimators consistently estimate $\alpha_1-\alpha_0$ in equation (2) under standard conditions. Parameters (3) and (5) are the same. In the context of estimating union-nonunion wage gaps, there is accumulating evidence that instrumental variables procedures "work" in the sense of producing approximately the same estimate of parameters (3) and (4) as are obtained from more complicated sample-selection-correction models. There is considerable evidence, however, that D is endogenous. See Duncan-Leigh and Robinson for evidence on these points.

These studies indicate that unobserved (by the econometrician) components of U_1-U_0 contribute negligibly to the endogeneity of D. Several economic models are consistent with this result. Uncertainty about U_1-U_0 at the time union membership decisions are made is one plausible story. Selection on the basis of nonwage components (nepotism, discrimination, etc.) is another source (see Robinson for an excellent discussion of this point). The most general case, that is econometrically the hardest, appears not to be the empirically fruitful one.

REFERENCES

Andrews, Donald, "Asymptotic Normality of Series Estimators For Nonparametric and Semiparametric Regression Models," Discussion Paper No. 874R, Cowles Foundation, Yale University, June 1989.

Chamberlain, Gary, "Asymptotic Efficiency in

Semiparametric Models With Censoring," *Journal of Econometrics*, Spring 1986, *32*, 189–218.

Cosslett, Stephen, "Semiparametric Estimation of a Regression Model with Sample Selectivity," in W. A. Barnett et al., eds., *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press, 1990.

Duncan, Gregory and Leigh, Duane, "The Endogeneity of Union Status: An Empirical Test," *Journal of Labor Economics*, July 1985, *3*, 385–402.

Gallant, Ronald and Nychka, Douglas, "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, March 1987, *55*, 363–90.

Heckman, James and Honoré, Bo, "The Empirical Content of The Roy Model," *Econometrica*, forthcoming 1990.

______ **and Robb, Richard,** "Alternative Methods for Evaluating The Impact of Treatment on Outcomes," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press, 1985.

Ichimura, Hidehiko and Lee, Lung Fei, "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in W. A. Barnett et al., eds., *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press, 1990.

Lee, Lung Fei, "Unionism and Relative Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review*, June 1978, *19*, 415–33.

Lewis, H. Gregg, *Union Relative Wage Effects: A Survey*, Chicago: University of Chicago Press, 1986.

Manski, Charles, "Identification of Binary Response Models," *Journal of the American Statistical Association*, September 1988, *83*, 729–38.

Newey, Whitney, "Two Step Series Estimation of Sample Selection Models," Department of Economics, Princeton University, October 1988.

Powell, James, "Semiparametric Estimation of Censored Selection Models," unpublished manuscript, University of Wisconsin-Madison, July 1989.

Robinson, Christopher, "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy*, June 1989, *97*, 639–67.

Nonparametric Bounds on Treatment Effects

By CHARLES F. MANSKI*

Assume that each member of a population is characterized by values for the variables (y_A, y_B, z, x). Here x is a vector describing a person and z is a binary variable indicating which of two treatments this person receives. The treatments are labelled A and B. The variables y_A and y_B are scalar measures of the outcomes of the two treatments.

For example, a cancer patient might be treated by (A) drug therapy or (B) surgery. The relevant outcome y might be life span following treatment. An unemployed worker might be given (A) vocational training or (B) job search assistance. Here the relevant outcome might be labor force status following treatment.

Assume that a random sample is drawn and that one observes the realizations of (z, x) and of the outcome under the treatment received. Thus y_A is observed if treatment A is received but is a latent variable if treatment B is received. Similarly, y_B is either observed or latent.

Suppose that one wants to learn the difference in expected outcome if all persons with attributes x were assigned to treatment A or B. This "treatment effect" is

$$
\begin{aligned}
(1)\quad t(x) &\equiv E(y_B|x) - E(y_A|x) \\
&= E(y_B|x, z = A)P(z = A|x) \\
&\quad + E(y_B|x, z = B)P(z = B|x) \\
&\quad - E(y_A|x, z = A)P(z = A|x) \\
&\quad - E(y_A|x, z = B)P(z = B|x).
\end{aligned}
$$

The central problem is identification. The data are from a population in which some people described by x received treatment A and the rest received B. The sampling process identifies the expected outcomes under the treatment received, $E(y_A|x, z = A)$ and $E(y_B|x, z = B)$. It also identifies the treatment-selection probabilities $P(z|x)$. But, the sampling process does not identify $E(y_A|x, z = B)$ and $E(z_B|x, z = A)$. Hence it does not identify the treatment effect.

An extensive literature on the estimation of treatment effects brings to bear prior information that, in conjunction with the sampling process, does identify $t(x)$. (See, for example, G. S. Maddala, 1983, and James Heckman and Richard Robb, 1985.) Two approaches have been dominant. One assumes that, conditional on x, y_A and y_B are mean-independent of z (i.e., $E(y_A|x, z = A) = E(y_A|x, z = B)$ and $E(y_B|x, z = A) = E(y_B|x, z = B)$). This assumption, routinely invoked in experiments with random assignment to treatment, implies that

$$
(2)\quad t(x) = E(y_B|x, z = B) - E(y_A|x, z = A).
$$

The second approach imposes identifying restrictions through a latent-variable model explaining treatment selections and outcomes. The latent-variable-model approach is widely used in settings where people self-select into treatment.

Suppose that one cannot confidently assert mean-independence, an identifying latent-variable model, or any other restriction that identifies the treatment effect. It might then seem that useful inference is impossible. This paper proves otherwise. Section I applies results from my earlier paper (1989) to show that an informative bound on the treatment effect holds whenever the outcomes y_A and y_B are themselves bounded. Section II derives a bound applicable when the treatment selection rule is to choose the treatment with the better outcome. Section III shows that the bounds of Sections I and II can be tightened if some component of x affects treatment selection but not treatment outcome. Section IV briefly discusses estimation of the bounds.

---

*Department of Economics and Institute for Research on Poverty, University of Wisconsin, Madison, WI 53706. This research is supported by NSF grant no. SES-8808276 and by a grant from the U.S. Department of Health and Human Services to the Institute for Research on Poverty. I have benefited from the comments of Arthur Goldberger and Robert Moffitt.

I. Bounded Outcomes

Suppose that, conditional on x, y_A is bounded within some known interval $[K_{A0x}, K_{A1x}]$, where $-\infty < K_{A0x} \le K_{A1x} < \infty$. Then obviously $E(y_A|x) \in [K_{A0x}, K_{A1x}]$. My earlier paper observes that the sampling process identifies a tighter bound, namely,

$$(3)\quad E(y_A|x) \in Y_A(x)$$

$$\equiv \big[E(y_A|x, z=A)P(z=A|x) + K_{A0x}P(z=B|x),$$

$$E(y_A|x, z=A)P(z=A|x) + K_{A1x}P(z=B|x)\big].$$

The lower bound is the value $E(y_A|x)$ takes if y_A equals its lower bound for all those who receive treatment B; the upper bound is determined similarly.

The width of the bound $Y_A(x)$ is $(K_{A1x} - K_{A0x})P(z=B|x)$. So the bound is informative if $P(z=B|x) < 1$. The bound is operational because $E(y_A|x, z=A)$ and $P(z|x)$ are identified by the sampling process. In practice, one can estimate $E(y_A|x, z=A)$ and $P(z|x)$ nonparametrically, yielding a nonparametric estimate for the bound (see Section IV).

Now suppose that y_B is also bounded, the interval being $[K_{B0x}, K_{B1x}]$. Then the treatment effect must lie in the interval $[K_{B0x} - K_{A1x}, K_{B1x} - K_{A0x}]$. The sampling process identifies a tighter bound, obtained by applying (3) and the analogous bound for $E(y_B|x)$ to (1). The result is

$$(4)\quad t(x) \in T(x)$$

$$\equiv \big[K_{B0x}P(z=A|x) + E(y_B|x, z=B)P(z=B|x) - E(y_A|x, z=A)P(z=A|x) - K_{A1x}P(z=B|x),$$

$$K_{B1x}P(z=A|x) + E(y_B|x, z=B)P(z=B|x) - E(y_A|x, z=A)P(z=A|x) - K_{A0x}P(z=B|x)\big].$$

The lower bound on $t(x)$ is the difference between the lower bound on $E(y_B|x)$ and the upper bound on $E(y_A|x)$. The upper bound on $t(x)$ is determined similarly.

The width of the bound $T(x)$ is

$$(5)\quad w(x) \equiv (K_{B1x} - K_{B0x})P(z=A|x) + (K_{A1x} - K_{A0x})P(z=B|x).$$

In general, this width depends on the treatment-selection probabilities $P(z|x)$. Suppose, however, that the bounds on y_A and y_B are the same (i.e., $[K_{A0x}, K_{A1x}] = [K_{B0x}, K_{B1x}]$). Then,

$$(6)\qquad w(x) = K_{1x} - K_{0x},$$

where $[K_{0x}, K_{1x}]$ is the common bound on the outcomes. The bound available without the sample data is $t(x) \in [K_{0x} - K_{1x}, K_{1x} - K_{0x}]$. Thus, when the bounds on y_A and y_B are the same, exploitation of the sampling process allows one to bound $t(x)$ to one-half of its otherwise possible range. In this case, the bound necessarily covers zero; it cannot identify the sign of the treatment effect.

The remainder of this section describes a class of applications in which the bound $T(x)$ is particularly useful—binary logical outcomes.

In many applications, the treatment outcome is a logical yes/no indicator, taking the value one or zero. For example, the outcome of a medical treatment may be (cured $=1$, not cured $=0$); the outcome of a vocational training program may be (completed $=1$, not completed $=0$). In both cases, $K_{A0x}=K_{B0x}=0$ and $K_{A1x}=K_{B1x}=1$, so the treatment effect must lie in the interval $[-1,1]$. The expected value of a one/zero indicator is the probability that the indicator equals one. So the bound $T(x)$ reduces to

$$
\begin{aligned}
(7)\quad & T(x) \\
&= \big[P(y_B=1|x,z=B)P(z=B|x) \\
&\quad -P(y_A=1|x,z=A)P(z=A|x) \\
&\quad -P(z=B|x),P(z=A|x) \\
&\quad +P(y_B=1|x,z=B)P(z=B|x) \\
&\quad -P(y_A=1|x,z=A)P(z=A|x)\big].
\end{aligned}
$$

The bound width is $w(x)=1$.

Binary logical variables are bounded by definition rather than by assumption. So we find that the sampling process alone, unaccompanied by prior information, suffices to bound the treatment effect to one-half its otherwise possible range.

II. Selection of The Treatment with the Better Outcome

In some settings the treatment-selection rule has the form

$$(8)\qquad z=B \Leftrightarrow y_B \geq y_A.$$

For example, a doctor may prescribe the more effective of two medical treatments. An unemployed worker may choose the employment program with the higher return.

If (8) holds, the bound $T(x)$ obtained in Section I can be tightened. By (8),

$$
\begin{aligned}
(9)\quad E(y_A|x,z=B) &= E(y_A|x,y_A\leq y_B) \\
&\leq E(y_A|x,y_A>y_B) \\
&= E(y_A|x,z=A) \\
E(y_B|x,z=A) &= E(y_B|x,y_B<y_A) \\
&\leq E(y_B|x,y_B\geq y_A) \\
&= E(y_B|x,z=B).
\end{aligned}
$$

Thus $E(y_A|x,z=A)$ and $E(y_B|x,z=B)$ are upper bounds on $E(y_A|x,z=B)$ and $E(y_B|x,z=A)$, respectively. The conditions $E(y_A|x,z=A)\leq K_{A1x}$ and $E(y_B|x,z=B)\leq K_{B1x}$ must hold. Hence knowing that (8) holds permits one to tighten the bound (4) on the treatment effect to

$$
\begin{aligned}
(10)\quad T(x) &= \big[K_{B0x}P(z=A|x) \\
&\quad +E(y_B|x,z=B)P(z=B|x) \\
&\quad -E(y_A|x,z=A),E(y_B|x,z=B) \\
&\quad -E(y_A|x,z=A)P(z=A|x) \\
&\quad -K_{A0x}P(z=B|x)\big].
\end{aligned}
$$

The tightened bound may or may not lie entirely to one side of zero. If it does, the sign of the treatment effect is identified.

III. Level-Set Restrictions

The bound $T(x)$ on the treatment effect at a given value of x does not constrain the treatment effect elsewhere. This is to be expected as no restrictions have been imposed on the behavior of $t(x)$ as a function of x. Suppose that one has information on the way $t(x)$, or its determinants, vary with x. Then one may be able to obtain a bound tighter than $T(x)$.

This section investigates the additional identifying power of level-set restrictions. A level-set restriction is an assertion that some function of x is constant on some $X_0 \subset X$, where X is the set of all possible values of x. An important special case is the exclusion restriction. Here one lets $x \equiv (x_1, x_2)$ and asserts that, holding x_1 fixed, a function of x does not vary with x_2. Thus the function is constant on the set $X_0 \equiv \{x_1\} \times X_2$, where $\{x_1\}$ is the set containing only the point x_1, and where X_2 is the set of all possible values of x_2.

A. *Level-Set Restrictions on the Treatment Effect*

It is often assumed in applications that the treatment effect does not vary with x. In particular, many studies specify a linear model with $E(y_A|x) = x\beta$ and $E(y_B|x) = x\beta + \alpha$, implying that $t(x) = \alpha$. The assumption that $t(x)$ is constant on all of X is a leading example of a level-set restriction.

Suppose it is known that $t(x)$ is constant on some set X_0. Then the collection of bounds $T(x)$, $x \in X_0$ must have a non-null intersection that contains the common value of the treatment effect. That is, for each $\xi \in X_0$,

$$(11)\quad t(\xi) \in T_0(X_0) \equiv \bigcap_{x \in X_0} T(x)$$
$$= \Big[\sup_{x \in X_0} \{ K_{B0x} P(z = A|x) + E(y_B|x, z = B) P(z = B|x) - E(y_A|x, z = A) P(z = A|x) - K_{A1x} P(z = B|x) \},$$
$$\inf_{x \in X_0} \{ K_{B1x} P(z = A|x) + E(y_B|x, z = B) P(z = B|x) - E(y_A|x, z = A) P(z = A|x) - K_{A0x} P(z = B|x) \} \Big].$$

The bound $T_0(X_0)$ improves on $T(\xi)$ for at least some ξ in X_0 unless $T(\cdot)$ is constant on X_0. Constancy of $T(\cdot)$ can occur in various ways. The one most likely to arise in practice is inclusion in x of an irrelevant component, one that affects neither K_{A0x}, K_{A1x}, K_{B0x}, K_{B1x}, $P(z|x)$, $E(y_A|x, z = A)$, nor $E(y_B|x, z = B)$. A restriction excluding this component from the treatment effect has no bite.

Although $T_0(X_0)$ improves on $T(x)$, it typically does not fix the sign of the treatment effect. If $T(x)$ covers zero for all $x \in X_0$, then so does $T_0(X_0)$. It was pointed out in Section I that $T(x)$ does cover zero whenever the bounds on y_A and y_B are the same.

B. *Level-Set Restrictions on the Outcome Regressions*

Suppose it is known that $E(y_A|x)$ is constant on some set $X_{A0} \subset X$, and that $E(y_B|x)$ is constant on some $X_{B0} \subset X$. (This includes cases in which one of the restrictions is trivial; the set X_{B0}, for example, might contain just one point.) Let $X_{AB0} \equiv X_{A0} \cap X_{B0}$. The reasoning used in Section A above implies that, for each $\xi \in X_{AB0}$, the bounds $Y_A(\xi)$ and $Y_B(\xi)$ defined in (3) can be tightened to

$$(12)\quad E(y_a|\xi) \in Y_{A0}(X_{A0}) \equiv \bigcap_{x \in X_{A0}} Y_A(x)$$
$$= \Big[\sup_{x \in X_{A0}} \{ E(y_A|x, z = A) P(z = A|x) + K_{A0x} P(z = B|x) \},$$
$$\inf_{x \in X_{A0}} \{ E(y_A|x, z = A) P(z = A|x + K_{A1x} P(z = B|x) \} \Big]$$
$$E(y_B|\xi) \in Y_{B0}(X_{B0}) \equiv \bigcap_{x \in X_{B0}} Y_B(x)$$
$$= \Big[\sup_{x \in X_{B0}} \{ E(y_B|x, z = B) P(z = B|x) + K_{B0x} P(z = A|x) \},$$
$$\inf_{x \in X_{B0}} \{ E(y_B|x, z = B) P(z = B|x) + K_{B1x} P(z = A|x) \} \Big].$$

These bounds on $E(y_A|x)$ and $E(y_B|x)$ imply a bound on $t(\xi)$, namely,

$$
\begin{aligned}
(13)\quad t(\xi) &\in T_{AB0}(X_{AB0}) \\
&\equiv \Big[\sup_{x\in X_{B0}} \{E(y_B|x,z=B)P(z=B|x) \\
&\quad + K_{B0x}P(z=A|x)\} \\
&\quad - \inf_{x\in X_{A0}} \{E(y_A|x,z=A) \\
&\quad \times P(z=A|x)+K_{A1x}P(z=B|x)\}, \\
&\inf_{x\in X_{B0}} \{E(y_B|x,z=B)P(z=B|x) \\
&\quad + K_{B1x}P(z=A|x)\} \\
&\quad - \sup_{x\in X_{A0}} \{E(y_A|x,z=A) \\
&\quad \times P(z=A|x)+K_{A0x}P(z=B|x)\}\Big].
\end{aligned}
$$

The treatment effect is constant on X_{AB0}. Hence the bound $T_0(X_{AB0})$ also applies here. Comparison of (11) and (13) shows that $T_{AB0}(X_{AB0}) \subset T_0(X_{AB0})$. It is intuitive that the present bound should improve on the earlier one. The derivation of $T_0(X_{AB0})$ presumed only that $t(x)$ is constant on X_{AB0}. The derivation of $T_{AB0}(X_{AB0})$ imposed the stronger restriction that $E(y_A|x)$ is constant on X_{A0} and $E(y_B|x)$ is constant on X_{B0}.

The bound $T_{AB0}(X_{AB0})$ may lie entirely to one side of zero. If so, the sign of the treatment effect is identified.

IV. Estimation of the Bounds

The bounds developed in Sections I, II, and III are functions of $E(y_A|x, z=A)$, $E(y_B|x, z=B)$, and $P(z|x)$. These quantities are identified by the sampling process and so generally can be estimated consistently.

If the conditioning variable x takes finitely many values, estimation is classical. $E(y_A|x, z=A)$, $E(y_B|x, z=B)$, and $P(z|x)$ can be estimated by the corresponding sample averages. $T(x)$ can be estimated by inserting these averages into (4) or (10), as appropriate. $T_0(X_0)$, being a finite intersection of the $T(x)$, can be estimated by the intersection of the estimates of $T(x)$. Assuming that the level-set restriction on $t(x)$ is correct, this intersection is nonempty with probability approaching one as the sample size increases. The bound $T_{AB0}(X_{AB0})$ can be estimated similarly.

If x has continuous components, nonparametric regression methods may be applied to estimate $E(y_A|x, z=A)$, $E(y_B|x, z=B)$, and $P(z|x)$. My earlier paper exposits the main issues and presents an empirical illustration estimating the bound $Y_A(x)$ on $E(y_A|x)$. This work can be applied directly to estimate the bound $T(x)$. Estimation of $T_0(X_0)$ is a more subtle problem, because this bound is the intersection of the infinitely many bounds $T(x)$, $x \in X_0$. A plausible approach, that warrants study, is to estimate $T_0(X_{N0})$. Here N is the sample size and X_{N0}, $N=1,\dots,\infty$ is a sequence of finite subsets of X_0, chosen to converge to a set dense in X_0. The bound $T_{AB0}(X_{AB0})$ could be estimated similarly.

REFERENCES

Heckman, James and Robb, Richard, "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press, 1985.

Manski, Charles F., "Anatomy of the Selection Problem," *Journal of Human Resources*, Summer 1989, *24*, 343–60.

Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Semiparametric Estimation of Selection Models: Some Empirical Results

By WHITNEY K. NEWEY, JAMES L. POWELL, AND JAMES R. WALKER*

Among the central theoretical developments in the econometric analysis of nonexperimental microeconomic data has been the analysis of selectivity bias. Following the work of James Heckman (1974), statistical techniques were developed in the 1970s to consistently estimate the parameters of these models. One potential drawback to the application of these techniques is their sensitivity to the assumed parametric distribution of the unobservable error terms in the model. In recent years, a number of estimation methods for selection models have been developed which do not impose parametric forms on error distributions; these methods are termed "semiparametric," since only part of the model of interest (the regression function) is parametrically specified.

While the statistical theory of these semiparametric estimators has received much attention, practical applications of the methods are lacking (an exception being the paper by Joel Horowitz and George Neumann, 1987). In this paper, we use semiparametric methods to reanalyze data on the labor supply of married women first studied by Thomas Mroz (1987) using parametric methods. The object of this reanalysis is to determine whether Mroz's results are sensitive to his parametric assumptions.

I. The Model and Estimators

To review the censored sample selection model considered here, the object of the statistical analysis is the estimation of the regression coefficients β in the equation

$$y = d \cdot [x'\beta + v], \tag{1}$$

where the dependent variable y is generated by the regressors x, unobservable error terms v, and an indicator variable d that denotes whether that dependent variable is uncensored. The indicator d is assumed to depend on a vector of conditioning variables w through a standard binary response model:

$$d = 1(w'\delta + \varepsilon > 0), \tag{2}$$

where $1(A)$ denotes an indicator function of the event A, δ is a vector of unknown coefficients, and ε is an unobserved error term. Dependence between the error terms v and ε in the "outcome" equation (1) and "selection" equation (2) introduces a nonlinear selection correction term into the model for the observed dependent variable; for observations with $d = 1$, the outcome equation (1) takes the form

$$y = x'\beta + \lambda(w'\delta) + u, \tag{3}$$

$$\lambda(w'\delta) \equiv E[v|\varepsilon > -w'\delta],$$

so that $E[u|w] = 0$. Validity of equation (3) requires that the original error terms v and ε be distributed independently of the conditioning variables w, or, more generally, that their conditional distribution given w depends only upon the "single index" $w'\delta$ appearing in the selection equation (2). Note that, since the components of x need not be included in the conditioning variables w, some regressors in equation (1) may be endogenous, as in the empirical application below.

A classical parametric approach to estimation of the parameters β and δ of this model would specify the joint distribution of the error terms v and ε as, say, bivariate normal, and then estimate these parameters along with the nuisance parameters of the assumed distribution by maximum likelihood. However, in practice, obtaining full maximum-likelihood estimates for equations (1) and (2) can be computationally cumbersome, as the joint likelihood may be ill-determined. The estimation scheme most frequently used in practice, proposed by Heckman, estimates the parameters in two stages. First, the parameters δ of the selection equation (2) are estimated using maximum likelihood, using the representation

---

*Newey: Princeton University and Bellcore; Powell and Walker: University of Wisconsin, Madison, WI 53706. This work was supported by the National Science Foundation, the Alfred P. Sloan Foundation, and the National Institute of Child Health and Development. We are grateful to Thomas Mroz for providing his data to us, to Arthur Goldberger for helpful comments, and to Kiho Jeong for computational assistance.

$$(4)\quad Pr\{d=1|w\} = E[d|w] = F(w'\delta),$$

where $F(\cdot)$ is the assumed cumulative distribution function of $-\varepsilon$ (typically standard normal). Then, given the estimator $\hat{\delta}$ from this first step, equation (3) is estimated by instrumental variables (classical least squares if x is a subset of w), substituting $\hat{\delta}$ for δ and using the parametric form of the correction function $\lambda(\cdot)$ (in the normal case, this is proportional to the inverse of the Mills ratio). Inference on the coefficients β can be carried out after adjusting the form of the asymptotic covariance matrix estimator of the instrumental variables estimator for preliminary estimation of the parameters δ in the selection equation.

If the joint distribution of the error terms in equations (1) and (2) is misspecified (either because the parametric form of the distribution is incorrect or because of neglected dependence of the conditional distribution on the index $w'\delta$) the second-step parameter estimator $\hat{\beta}$ will be inconsistent in general. To obtain consistent estimators of these parameters when the joint distribution of the error terms is of unknown form (assuming the forms of the regression functions in (3) and (4) are correctly specified), the same two-step estimation strategy can be adopted, and at the same time the parametric restrictions imposed on the functions $F(\cdot)$ and $\lambda(\cdot)$ are relaxed. Several estimators of δ in equation (4) have been proposed in the recent literature on semiparametric estimation; while some of these methods sidestep estimation of the function $F(\cdot)$, the methods we consider here use nonparametric regression methods to estimate $F(\cdot)$ along with the parameters δ. Similarly, the methods for estimation of the parameters β in the second stage also involve nonparametric regression methods, that are applied either to estimation of $\lambda(\cdot)$ directly or to estimation of other regression functions which depend upon the index $w'\delta$.

The two estimators of δ employed below were proposed by Hidehiko Ichimura (1987), and by Roger Klein and Richard Spady (1987), respectively; the former chooses $\hat{\delta}$ to minimize a least-squares criterion, while the latter chooses $\hat{\delta}$ to maximize a quasi-likelihood function. In both cases, the nonparametric estimator of $F(w_i'\delta)$ for observation i is a kernel estimator, taking the form of a weighted average of the indicator variables d_j across the index j with weights given by $K[(w_i - w_j)'\hat{\delta}/h]$, for $K[\cdot]$ a "kernel" function that tends to zero as the magnitude of its argument increases. For this estimator, the "bandwidth" parameter h governs the degree of "smoothness" imposed on the estimated function $\hat{F}(\cdot)$, with larger values of h corresponding to a smoother function estimate. Ichimura and Klein-Spady prove that their respective estimators of δ are consistent (up to a normalization on δ) and asymptotically normal under suitable conditions on the model, kernel function, and bandwidth.

Given an estimator $\hat{\delta}$ of the parameters in the first step, two approaches to estimation of β in equation (3) have recently been considered. One approach follows Peter Robinson (1988) in using (3) and the identity $\lambda(w'\delta) \equiv E[\lambda(w'\delta)|w'\delta]$ to obtain

$$(5)\quad y - E[y|w'\delta] = (x - E[x|w'\delta])'\beta + u.$$

Estimators of the conditional expectation functions in this equation can be obtained using kernel estimators of the same form described above, replacing d with y and x. Then, with the corresponding estimators of the dependent and explanatory variables in equation (10), the coefficient vector β can be estimated using two-stage least squares, with functions of w used as instrumental variables. Powell (1987) analyzes the behavior of this estimator with instrumental variables of the form $\hat{f} \cdot z$, where z is a vector of functions of w with the same dimension as x and $\hat{f}$ is a nonparametric estimate of the density function of $w'\delta$ (which is proportional to the denominator of the kernel estimators). Under conditions on the model, kernel function, bandwidth, and preliminary estimator $\hat{\delta}$, Powell shows that the resulting estimator $\hat{\beta}$ is consistent and asymptotically normal, and gives estimators of its covariance matrix which account for the first-step estimates.

The foregoing estimator essentially sidesteps estimation of the "nuisance" function $\lambda(\cdot)$ in equation (3) (although an estimator can easily be constructed). Another approach to estimation of equation (3) simultaneously estimates a "series approximation" to $\lambda(\cdot)$. That is, by making the approximation

$$\lambda(w'\hat{\delta}) \cong \sum_{j=1}^{J} \eta_j \cdot p_j(w'\hat{\delta}), \tag{6}$$

where the $\{\eta_j\}$ are unknown coefficients and the functions $\{p_j(\cdot)\}$ are known "basis" functions; substitution of (6) into equation (3) yields a model that is linear in parameters, and thus estimable by standard methods. The number of basis functions, J, serves a similar role here to the bandwidth parameter h for kernel estimation—by increasing J, the approximating function becomes more flexible, but at the cost of introducing more nuisance parameters to estimate. This approach was taken by Lung-Fei Lee (1982) and Stephen Cosslett (1984), among others; Newey (1988) gives conditions on the model, basis functions, and number of terms under which the resulting estimator of β is consistent and asymptotically normal, and has suggested a consistent estimator of its asymptotic covariance matrix.

All of the estimators above require some mechanism to choose the amount of "smoothing" imposed (either the bandwidth or number of basis functions). While theoretical conditions are available relating the rate of change of the smoothing parameter to the increase in the sample size, such conditions are of little guidance for a particular sample. For the results below, the smoothing parameter is chosen by a method developed for nonparametric regression by P. Craven and Grace Wahba (1979), termed "generalized cross-validation." This approach is adopted here in the hope that it will perform well in this semiparametric setting, as it is known to do for nonparametric regression.

II. The Data and Estimates

Mroz's study of the labor supply of married women is a leading example of the application of parametric techniques to selectivity bias models. The data set Mroz analyzed consists of measurements on characteristics of 753 married women, of whom 428 were working at the time of the study. The dependent variable y is the annual hours of work; the regressors x for this equation included the logarithm of the wage rate (assumed endogenous), family income less wife's labor income (in thousands of dollars), indicators for young and older children in the family, and the wife's number of years of age and education. Mroz considered several variations on this model; in the version we reanalyze here (summarized in his Table X), the conditioning variables w included the exogenous variables above plus years of labor force experience, other background variables, and various interaction terms (a more complete description of the data set is given in his paper). Mroz specified several different parametric models for the disturbance terms in equations (1) and (2), and found that the point estimates of the coefficients in β were similar across specifications; here, we check whether this result is specific to the particular parametric models he chose.

In the first stage of our analysis, we calculated the Ichimura and Klein-Spady estimators using labor force participation as the variable d, and the same 18-variable specification of w used in Mroz's Table X; for comparison purposes, we also calculated the probit maximum likelihood estimator. For the semiparametric estimates, the kernel function $K(\cdot)$ was taken as a standard normal density function; the bandwidth h, determined by generalized cross validation, was

TABLE 1—ESTIMATES OF HOURS EQUATION PARAMETERS

Variable	(1)	(2)	(3)	(4)
Log Wage	672	122	167	64
	(217)	(219)	(341)	(253)
Nonwife	−6.5	3.8	6.3	4.4
Income	(3.6)	(4.6)	(5.6)	(4.9)
Young	−284	55	70	76
Children	(147)	(174)	(203)	(182)
Older	−85	−87	−79	−88
Children	(36)	(35)	(37)	(37)
Age	−9.1	3.4	8.1	4.4
	(7.0)	(7.1)	(8.0)	(7.6)
Education	−86	−86	−88	−86
	(29)	(30)	(29)	(31)

Note: Col. (1) Two-stage least squares; Col. (2) Normal two-step; Col. (3) Weighted kernel; Col. (4) Series.

TABLE 2—*Chi*-SQUARE TESTS OF SPECIFICATION

Difference	Statistic	Degrees of Freedom	*P*-Value
(1) vs. (2)	11.26	1	0.001
(1) vs. (3)	10.76	6	0.094
(1) vs. (4)	10.49	2	0.005
(2) vs. (3)	1.10	6	0.982
(2) vs. (4)	0.40	1	0.527

set at 0.4 times the standard deviation of the fitted index $w'\hat{\delta}$. The coefficient estimates for the semiparametric methods turned out to be quite close to the probit estimates; the estimated standard errors for the semiparametric estimators were slightly larger than for probit, with increases of 1 to 25 percent (with a median of 14 percent) for Ichimura's estimator and increases of 2 to 24 percent (with a median of 13 percent) for the Klein-Spady estimator.

A formal test for normal disturbances can be based upon the normalized difference between the probit and semiparametric estimates; the test statistics for this hypothesis, which have asymptotic *chi*-squared distributions with 17 degrees of freedom if the probit model is correct, were calculated to be 8.9 for Ichimura's estimator and 7.1 for the Klein-Spady estimator. These values are far below standard *chi*-square critical values, so that the hypothesis of normal errors in equation (2) cannot be rejected; as a result, we take the first-step estimator $\hat{\delta}$ to be the probit maximum-likelihood estimator when computing the second-step estimates below.

Results of the second-step estimates of the coefficients β are given in Table 1, along with their standard error estimates. For comparison purposes, two-stage least squares estimates (without selection correction) and Heckman two-step estimates assuming normal errors are reported; due to differences in computational algorithms and standard error estimators, these differ somewhat from the values reported in Mroz's Table X. Generalized cross validation yielded a value of the bandwidth h of 0.1 times the standard deviation of the first-step index $w'\hat{\delta}$ for the weighted kernel estimator, where again a normal kernel was used. The basis functions for the series estimator were taken to be the inverse Mills' ratio multiplying powers of the first-step index, with number of terms (chosen by generalized cross validation) J equal to 2 (i.e., a single additional term beyond the normal selection correction term).

As the entries in Table 1 indicate, the semiparametric estimators give results that are very similar to the normal two-step estimators, and different from the two-stage least squares estimates. Formal test statistics for the null hypothesis of no difference in the probability limits of the estimators are tabulated in Table 2; these statistics are calculated as quadratic forms of the differences in the six reported coefficient estimates about the inverses of the asymptotic covariance matrix estimates of these difference. The test statistics that compare the normal two-step estimator and the series estimator to two-stage least squares both give strong evidence of the presence of selectivity bias, as their marginal significance levels are much smaller than conventional values. The corresponding test statistic for the weighted kernel estimator, while comparable in magnitude to the others, has a higher marginal significance level, due to the increase in the degrees of freedom for this test. Finally, the last two lines of Table 2 indicate that the semiparametric estimators do not give significantly different results from the two-step estimator based upon the assumption of jointly normal errors.

III. Conclusions

The results given in the previous section indicate that application of the semiparametric estimation methods to this data set give estimates similar to parametric methods assuming normal errors, but provide somewhat weaker evidence of the presence of selectivity bias. These results are not entirely surprising, since Mroz computed analogous estimates under two other nonnormal parametric models for the error distribution, and also found them to be similar to the Gaussian estimates. In contrast, Mroz found that parameter estimates varied widely across different specifications of the vector of regressors x and conditioning variables w; the results reported here can be viewed as further evidence that specification of the regression function and set of instrumental variables appears to be more important than specification of the error distribution for these data.

REFERENCES

Cosslett, Stephen R., "Semiparametric Estimation of a Regression Model with Sample Selectivity," University of Florida, June 1984.

Craven, P. and Wahba, Grace, "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, January 1979, *31*, 377–403.

Heckman, James J., "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, July 1974, *42*, 679–93.

Horowitz, Joel L. and Neumann, George R., "Semiparametric Estimation of Employment Duration Models," *Econometric Reviews*, No. 1, 1987, *6*, 1–40.

Ichimura, Hidehiko, "Consistent Estimation of Index Model Coefficients," unpublished doctoral dissertation, MIT, June 1987.

Klein, Roger W. and Spady, Richard S., "An Efficient Semiparametric Estimator of the Binary Response Model," Bellcore, June 1987.

Lee, Lung-Fei, "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, July 1982, *49*, 355–72.

Mroz, Thomas A., "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, July 1987, *55*, 765–99.

Newey, Whitney K., "Two Step Series Estimation of Sample Selection Models," Princeton University, April 1988.

Powell, James L., "Semiparametric Estimation of Bivariate Latent Variable Models," Working Paper No. 8704, SSRI, University of Wisconsin-Madison, July 1987.

Robinson, Peter M., "Root-N-Consistent Semiparametric Regression," *Econometrica*, July 1988, *56*, 931–54.

Reviving the Federal Statistical System: The View From Academia

By JEFFREY A. MIRON AND CHRISTINA D. ROMER*

There is a tendency to think of official government statistics as unambiguous measures of economic activity. In truth, however, nearly all of the available series are based heavily on assumptions, and are sensitive to the estimation techniques used. Furthermore, many of these assumptions and estimation techniques have been refined and improved over time. Whether these underlying assumptions are reasonable and whether the refinement of assumptions over time has been useful depends crucially on the questions one is trying to answer. For example, an estimate of consumption derived from data on retail sales may be perfectly adequate for planning future production, or setting government budgets, but may be disastrous for testing a subtle economic theory. Similarly, gathering more genuine consumption data might improve our current estimates of consumption, but a series that reflects retail sales for one era and genuine consumption for another could wreak havoc when used in estimating a time-series relationship.

Given that the adequacy of various statistics depends crucially on the use one wishes to make of them, reviving the federal statistical system is clearly a delicate proposition that must try to balance the needs of many different users. In this paper, we discuss some of the concerns about the government statistics that are most relevant to academic researchers and suggest revision procedures that would address some of these concerns.

I. Assumptions

One issue of concern to academic users is the role of assumptions in government statistics. While it is inevitable that some assumptions must be used to derive a quantitative portrait of the very complex U.S. economy, such assumptions can differ tremendously in their reasonableness and in their effect on research studies.

An example of a series where assumptions play a crucial role is the Federal Reserve Board (FRB) index of industrial production. The FRB index is estimated using data on labor input and electricity consumption in addition to data on actual physical production. In 1977, for example, only 42.9 percent of the value-added represented by the index was estimated directly from data on physical units of output, while 30 percent was estimated from data on kilowatt-hours and 25.2 percent from data on production-worker-hours (see FRB, 1986, p. 34). To use the data on labor input to estimate production, the FRB makes two crucial assumptions about the behavior of productivity. First, the FRB makes an adjustment for the cyclical behavior of productivity. Second, the FRB does not make any further adjustment for the seasonal behavior of productivity; that is, it assumes that there are no short-run increasing returns to labor over the seasonal cycle.

The possible biases or problems introduced by these assumptions are many. Most obviously, the use of labor input in the FRB index means that many components of the index will not provide independent evidence on the cyclical behavior of labor productiv-

---

[†] *Discussants*: F. Thomas Juster, University of Michigan; Richard Ruggles, Yale University; James R. Hines, Jr., Princeton University.

*Department of Economics, Boston University, Boston, MA 02215, and Department of Economics, University of California, Berkeley, CA 94720, respectively.

ity; any regression of output on labor input will partly recover the FRB's assumptions about productivity. The asymmetry in the FRB's treatment of seasonal and nonseasonal variation in productivity means that the monthly FRB index may provide flawed comparisons of seasonal cycles and business cycles. Evidence that this is true is provided by J. Joseph Beaulieu and Miron (1989) who find that, for industries for which physical production data are available, productivity varies significantly over the seasonal cycle.

Another example of an official series where assumptions may lead researchers astray is the consumer expenditures series. A recent study by David Wilcox (1989) discusses the fact that several components of total consumer expenditures, particularly food, clothing, and furniture, are estimated from data on retail sales. The estimation of consumption from retail sales data is complicated by the fact that the sales data are collected by type of store, rather than by type of product. To estimate the consumption of various products from these sales data, the Bureau of Economic Analysis (BEA) assumes that the composition of sales at general merchandise stores, such as department stores, is invariant to the state of the business cycle and the time of year.

Wilcox provides evidence that this assumption may be at odds with actual behavior. For example, he shows that the ratio of sales at furniture stores to sales at apparel stores is highly variable both seasonally and cyclically. As a result, it is unlikely that furniture sales are proportional to apparel sales within department stores. The effects of this faulty assumption may be quite large because department stores account for a large fraction of total sales. For example, Wilcox suggests that the procedures may cause the estimated consumption of durable goods to be more highly correlated with consumption of nondurables than is actually true. This bias could certainly distort studies of the relative behavior of different types of goods.

These examples of the problems posed for academic researchers by assumptions in the government statistics point out some obvious suggestions for reviving the federal statistical system. First, priority should be given to deriving series based on more actual data and on fewer assumptions. For example, it might be desirable to try to develop direct measures of physical production for the categories of industrial production that are now measured exclusively by labor or energy input. A related change would be to conduct more frequent surveys or experiments that could be used to test or modify the assumptions underlying the high frequency series.

Second, a more modest suggestion would be to revise the way that the government statistics are made available to researchers. One major improvement would be to provide both the final series based on various assumptions and the actual base data. This way researchers could better gauge the importance of the assumptions and could opt to use the data in a less adulterated form if that were more appropriate. A related change in presentation would be to state the assumptions more clearly. Many researchers are unaware of the assumptions that are used and tracking down the assumptions is often very difficult. Thus, more resources should be devoted to preparing detailed descriptions of the exact procedures used to derive the official statistics.

A particular case where making the data available in an unadulterated form is crucially important involves seasonal adjustment. While many of the government statistics are produced on both an adjusted and unadjusted basis, one of the most important and widely used set of series, the real quantities from the *National Income and Product Accounts*, is only available in seasonally adjusted form (see BEA, 1986).

For academic researchers, seasonally adjusted data present several problems. First, the Census X-11 procedure used by most government agencies yields data that are inappropriate for testing economic models based on rational expectations. This is true because these procedures make current data a function of future observations. More importantly, seasonally adjusting the data may remove some of the most significant fluctuations in the series. Recent studies by Robert Barsky and Miron (1989) and Beaulieu-Miron indicate that seasonal variation in aggregate series such as industrial production accounts for a very large fraction of the total variation in these series. As a result, it is important to have unadjusted data available so that researchers can try to better understand the source of such seasonal fluctuations and the relation of seasonal fluctuations to cyclical movements.

II. Consistency

Another characteristic of economic statistics that is particularly important to academic users is consistency over time. Much economic research and especially much historical research seeks to identify changes in economic behavior over time. Even studies that do not explicitly compare different time periods often make use of long time-series. This is especially true of macroeconomic analyses, where it is necessary to use data spanning at least several decades to have enough observations from which to draw precise inferences. Given that many of the questions of interest to economists can only be answered using long time-series, comparability or consistency of the data over time is crucial. If the methods or concepts used to derive the statistics change substantially, then comparisons of different eras become meaningless and time-series regressions become very hard to interpret.

The statistics covering the longest time periods are, unfortunately, the ones most likely to have problems of consistency. Recent research indicates that several of the long time-series are constructed in very different ways in the prewar and postwar eras. An example of an inconsistent series is the aggregate unemployment series published by the BEA in its compendium *Long Term Economic Growth*. Whereas the data since 1940 are based on a contemporaneous survey of households, the data for the period 1900–30 were constructed retrospectively using data from the decennial *Census of Population* and fragments of employment and output data. Romer (1986) shows that the assumptions used to convert the available data into estimates of the unemployment rate systematically overstate the swings in unemployment in the prewar era. The sources of this bias are the fact that changes in production are often used to proxy for changes in employment, without any correction for the cyclical behavior of productivity, and the fact that the labor force is assumed to have no cyclical variation. Because the early unemployment estimates are, by construction, more volatile than the later data, this series will overstate the decline in the severity of business cycles that has occurred over time.

The implications of the inconsistencies between prewar and postwar statistics for reviving the federal statistical system are many. One obvious one is that additional resources and effort might be used to improve the existing prewar estimates of many economic aggregates. While nothing can undo the fact that certain crucial surveys were not undertaken before World War II, the advent of computers and the creative use of previously neglected data for certain states or for a cross section of cities could lead to prewar series that are substantially more consistent with modern data than are the currently published prewar statistics.

A more modest approach to the incomparability of certain series over time would be to document the existing prewar series more thoroughly and to flag the inconsistencies more diligently. An example of where such documentation would be very useful is for the estimates of GNP for 1909–28. As discussed in Romer (1988), the *National Income and Product Accounts* reproduces this series but fails to emphasize that it is derived using data and methods that are very different from modern estimates. As a result, many researchers have incorrectly inferred that the pre-1929 series is consistent with modern GNP data.

While inconsistencies are most profound in the longer time series, they also exist in many of the government statistics for the postwar era. An example of such an inconsistency occurs in the labor force and unemployment data for the postwar era. Beginning in 1967 the questions used to establish unemployment were tightened so that, among other things, inactive job seekers were excluded from the unemployment statistics. According to a study by Robert Stein (1967), the effects of this change were substantial. For example, in 1966, the only year of overlap between the old and new procedures, the change in questions reduced the unemployment rate for adult men from 2.5 percent to 2.2 percent. Clearly, such a change in the series could have a significant impact on studies of the cyclical behavior of the labor force or unemployment.

The problems created for researchers by revisions in procedures and definitions suggest several guidelines for future revisions of existing postwar series. One obvious guideline is that data revisions should not be undertaken lightly. It is often easy to think of a slightly better way to ask a question or to define a quantity. However, given that even slight changes in procedures can have large effects, the decline in consistency could make such marginal improvements undesirable for academic users. A related point is that if one is going to revise a series, it should be done all at once, rather than gradually over time. A single major break in the series is easier to flag and impress upon users than a series of minor changes made at different times. This suggests that a possible strategy for reviving the federal statistical system would be to completely overhaul one or two series a year, rather than to make modest changes in many series at once.

Another guideline suggested by a desire for consistency is that revisions should be taken back as far as possible. In some cases, such as a change in the questions asked on a survey, a retrospective adjustment is not possible. In other cases, it is possible to at least make a reasonable adjustment going back in time. Indeed, this has been done recently with the price index for computers used in the derivation of the GNP series. A hedonic index was derived going back to the late 1960s so that computers could be deflated on a consistent basis since their earliest widespread use (see David Cartwright, 1986). Such retroactive adjustments are very useful and should have a high priority.

A final guideline is that when revisions cannot be taken back in time, it is crucial to provide a period of overlap between the old and new procedures. Such a period of overlap allows researchers to measure precisely the effects of the change in statistical methods. Moreover, the period of overlap should be fairly long: there should be comparison observations from different points in the business cycle so that researchers can test whether the revisions affect only the level of the series or also its cyclical properties. In the case of the 1967 revision of the unemployment statistics, the period of overlap was only one year long. This is unfortunate because it seems very likely that changing the treatment of discouraged workers could have a very different effect in recessions than in booms.

REFERENCES

Barsky, Robert B. and Miron, Jeffrey A., "The Seasonal Cycle and the Business Cycle," *Journal of Political Economy*, June 1989, *97*, 503–34.

Beaulieu, J. Joseph and Miron, Jeffrey A., "Seasonality in U.S. Manufacturing," unpublished manuscript, Boston University, 1989.

Cartwright, David W., "Improved Deflation of Purchases of Computers," *Survey of Current Business*, March 1986, *66*, 7–10.

Romer, Christina D., "Spurious Volatility in Historical Unemployment Data," *Journal of Political Economy*, February 1986, *94*, 1–37.

______, "World War I and the Postwar Depression: A Reinterpretation Based on Alternative Estimates of GNP," *Journal of Monetary Economics*, July 1988, *22*, 91–115.

Stein, Robert L., "New Definitions for Employment and Unemployment," *Employment and Earnings and Monthly Report on the Labor Force*, February 1967, *13*, 3–27.

Wilcox, David, "What Do We Know About Consumption?," unpublished manuscript, Board of Governors of the Federal Reserve System, 1989.

U.S. Board of Governors of the Federal Reserve System, *Industrial Production*, 1986 Edition, Washington: USGPO, 1986.

U.S. Bureau of Economic Analysis, *Long Term Economic Growth*, Washington: USGPO, 1973.

______, *The National Income and Product Accounts of the United States, 1929–82*, Washington: USGPO, 1986.

Reviving the Federal Statistical System: A View from Industry

By ROSANNE COLE*

The view expressed in this paper is taken from the perspective of a data user, employed by a firm that, like all other business firms in the United States, provides data to the federal statistical system. Some problems with the data are described first. They range from reporter error to errors arising from statistical practices. Suggestions for improving the system are then discussed. Particular attention is given to the Juster report.

I. Examples of Problems with Federal Statistics

My own and my colleagues' experience provide some examples of nonsampling error and how they arise. Such problems may occur, not when the questions asked are unclear—for then a respondent is quick to ask the statistical agency for clarification. They more often occur when a question seems clear, but is misunderstood because the respondent does not know the reason for the question or how the answer will be used.

Consider the following example. An agency produces a monthly series corrected for seasonal variation with historical seasonal factors. The data source on which some reporters rely, however, may not contain data referring to a calendar month. In order to cope with this complication, the reporting form contains a "number of weeks" question.

Now suppose a reporter changed from a data source that contained some 2- and 6-week months to a source that contained calendar week months. Suppose further that the week's question was understood by the respondent to mean not the number of weeks to which the data reported refer, but the number of weeks in the month. There is a good chance that the published series would show some very peculiar movements. One can only conjecture how common that simple error may be, and whether it is a reason why concurrent seasonal adjustment sometimes makes such a difference.

A second example illustrates the confusion created by the continued ambiguities in the SIC and of plants not asked the right questions in the 1987 *Census of Manufactures*. Data on materials consumed were needed by my colleague, Ellen Dulberger, in an effort to develop improved measures of value-added and productivity in the computing equipment industry. The logic and main memory of computers is packaged on boards or in complex electronic modules. Imagine her surprise to find these products not shown as consumed by the computing equipment industry in the preliminary tabulation of the 1987 Census data.

A walk down the hall confirmed the bad news and produced even more. Plants producing computers were asked if they consumed ball bearings or paperboard boxes, but not if they consumed boards or complex modules. One reason for the failure to ask about these products is that the Census, though tabulated on the newly revised SIC basis, was conducted on the old SIC basis. But even on the basis of the newly revised SIC, they could still equally well be coded computer parts as electronic components, unless a respondent concludes from the choices given that "other materials and components" is a better descriptor.

Plants producing these computer components were not asked if they consumed semiconductor chips. On the other hand, plants producing chips were asked if they consumed computer parts.

---

*IBM Corporation, Armonk, NY 10504. I am indebted to my colleague, Ellen Dulberger, for valuable comments and for bringing to my attention many of the data problems cited herein. Conversations with colleagues Joseph J. Smith and Joseph F. Scannello have given me an appreciation of the effort it takes for knowledgeable professionals to report data for a large firm. Interaction with personnel from the mainline statistical agencies has given me an appreciation of their responsiveness to problems brought to their attention and of the efforts they make to get the data right, useful, and relevant. In important respects, the problems with the federal statistical system are not with these examples of its human capital.

That is an extremely simplified characterization of a very complicated set of problems. In a narrow, technical sense, the forms were correct and the respondents filled them out correctly. But, what is the value of the data generated?

Semiconductor manufacturing equipment is lumped into a 4-digit "not elsewhere classified" category, along with broom-making machinery, buttonhole and eyelet machines, cotton ginning machines, cement making, hat making, and other miscellaneous, specialized machinery. The 1987 SIC revision carved out twenty-six categories of semiconductor manufacturing equipment within that 4-digit collection. This comes a bit late, for this equipment seems destined never to appear in federal statistics. As the press has observed (*New York Times*, November 27, 1989, p. 1), not much of it is still produced in the United States, and the new Harmonized System for trade data contains no classification for these $1 to $2 million dollar tools.

Software products are classified in the service sector as part of business services. For close to a decade, one has been able to observe that so-called service purchased in small packages by consumers from computer stores or complex operating systems arriving at business sites in trucks. Blank storage media are treated as goods, but when they contain coded information, are regarded as services by the SIC. This present odd distinction between goods and services will perhaps vanish in the future as software products are delivered electronically to users.

The effort to develop improved estimates of value-added was motivated by some old problems that became visible and important when the new price index for computing equipment was introduced into the *National Income and Product Accounts* with the 1985 benchmark revision. This price index is explicitly corrected for quality changes and shows price declines averaging about 15 percent per year. The price indexes available for deflating the industry's consumption of materials (produced here and abroad) are not. This inconsistency could create an overestimate of "real" value-added growth in the computing equipment industry and a corresponding underestimate elsewhere: partly in real value-added growth in domestic materials producing industries and partly in the real growth of imported materials.

The Producer Price Index for semiconductors is a case in point. This index does not seem credible to those familiar with these products and their prices. Whether the problem is one of the sample of products priced, one of index construction procedures, or both, is not known. Some preliminary research by Dulberger (1989) suggests that the answer may be both.

Another old problem that came to light is the "index number problem." When there are large changes in relative prices, estimates of growth in real values of aggregates become more sensitive to the choice of the base year because of substitution bias. For example, real growth in gross product originating in manufacturing (expressed in 1982 prices) is much larger than what real growth in manufacturing (expressed in 1989 prices) would be because relative prices of computing equipment are substantially lower in 1989 than they were in 1982. The same can be said for estimates of real growth in the nation's GNP, in its Producer Durable Equipment (PDE) component, and in the office and computing equipment component of PDE.

The magnitude of this bias can be important, as the Bureau of Economic Analysis (BEA) (1988) and Jack Triplett (1989) have pointed out. One would conjecture that substitution bias may have played some role in the construction sector productivity mystery inasmuch as this sector's gross product, expressed in 1972 prices, grew by 1.4 percent per year, but when expressed in 1982 prices, fell by 1.9 percent per year over the 1966–73 period. Many users look forward to the abandoning of the present fixed-weight index procedures.

As a user, I am particularly interested in following price, employment, domestic production, and trade data for the information systems industry. Employment is the only one of these measures available on a timely monthly basis for the software, services, and equipment segments of the industry, thus enabling the construction of an aggregate.

Recent revisions of the employment statistics have been unusually large—on the order of 6 to 7 percent of the number employed in 1988. The BLS practice and procedure with respect to revisions is even more troubling. Specifically, it is hard to understand why the (790) monthly sample survey data are not used to supplement the broader-based (ES-202) benchmark data instead of the other way round. The benchmark data are considered more reliable, are released quarterly, and contain monthly information. Yet only a small part of this preferable monthly information is ever reflected in the monthly statistics; the majority is essentially discarded.

The preceding examples are of traditional problems, the resolution of which could go far toward removing some of the present inconsistencies in, and perceived irrelevance of, federal statistics. The fundamental problem is one of making the federal statistical system more responsive to emerging data needs.

II. Strengthening the Federal System

The Juster report was originally the preliminary report of the AEA Committee on the Quality of Economic Statistics (1988). Prepared by the chairman, F. Thomas Juster, it dealt with the question of what needs to be done to revive the system, and recognized that the answer is not simply a matter of additional resources.

The report emphasized two changes that are, in the minds of the committee, critical to the long-run viability of the system: the establishment of 1) an effective statistical policy function within the Office of Management and Budget; and 2) active research and development programs within the mainline statistical agencies. A coherent, consistent body of economic information is unlikely to be produced by our decentralized federal statistical system without an effective central coordinating office. And "effective" should be understood to mean an office concerned mainly with standards for statistical programs and products, charged with a priority-setting function, and possessing sufficient professional competence, stature, and budgetary clout to carry out its responsibilities.

The Juster report considers establishing internal research capabilities within the statistical agencies the most important single factor for maintaining the quality and relevance of federal statistics. It cautions, however, that an active internal program of research and development requires a number of conditions that are not presently in place. For example, there needs to be steady pressure from a statistical standards office; there needs to be provision of adequate resources from which there may be little short-term payoff; and equally important, the economics profession itself needs to provide the training required to produce the skilled personnel with which to staff a statistical research and development function.

Two questions naturally arise with respect to the Juster report recommendations. Are these changes sufficient, and what are their prospects? Even though an internal research capacity is critical to maintaining the vitality of federal statistical programs, it will not do so unless the research output is effectively linked to and reflected in an agency's ongoing programs. A research group functioning merely as an ornamental appendage is both wasteful and a source of frustration to its members. The concern here is not the need for, or relevance of, a research activity; it is with the potential for isolation of ongoing programs from that activity.

In many respects, and as the Juster report implies, the prospects for these changes are presently not very bright. The changes are certainly needed. But efforts to institute them are sure to face the severe constraints of limited statistical budgets and a continued emphasis on reduction of paperwork burden.

The prospects would be somewhat improved if the effort gained the support of the business community. This support is essential to the success of many new measurement efforts inasmuch as business firms are a primary source of the raw data. Business firms are devoting as much effort to reducing internal as they are to reducing external paperwork burden. Support of the business community is likely to be stronger, therefore, the greater the chance that the data required for new measurements already reside in current business measurement and accounting systems.

Most business executives are aware of the force with which changes in aggregate economic conditions affect their business. They are also aware that government policy actions can be affected by measures of these aggregates. Accordingly, there is generally strong support for improving various macroeconomic indicators. There is somewhat less appreciation, however, for the micro foundations of these statistics. Here, in my opinion, business economists can perform a unique service within their firms, both as teachers and as advocates for improved measurements as well as for improved reporting of data.

Our system of economic intelligence has undergone a prolonged period of budgetary constraint. During that time, the base has been steadily depreciating with only patches applied to meet the increased demands on it. As with any neglected asset, it will take resources to repair. And it will also require an unusual amount of cooperation and support from the private sector.

REFERENCES

Dulberger, Ellen, "Measurement and Use of Alternative Price Indexes for Selected Semiconductors," paper prepared for NBER Price and Output Measurement Workshop, July 1989.

Juster, F. Thomas, "The State of U.S. Economic Statistics: Current and Prospective Quality, Policy Needs, and Resources," paper prepared for 50th Anniversary Conference, NBER Conference on Research in Income and Wealth, May 1988.

Triplett, Jack E., "Superlative and Quasi-Superlative Indexes of Price and Output for Investment Goods: Office, Computing, and Accounting Machinery," paper prepared for NBER Price and Output Measurement Workshop, July 1989.

U.S. Bureau of Economic Analysis, "Gross Product by Industry: Comments on Recent Criticisms," *Survey of Current Business*, July 1988, 132–33.

Reviving the Federal Statistical System: International Aspects

By ROBERT E. LIPSEY*

Data on the international economic relations of the United States are collected by, and processed in, several different government agencies. I think of the following as the major elements of U.S. international data:

1) Merchandise trade values, collected by the U.S. Customs Service and processed by the Bureau of the Census;

2) Trade in services, direct investment flows and stocks, and the balance of payments, all the responsibility of the Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce;

3) Prices in international trade transactions, of interest by themselves and also the basis for the deflation of merchandise trade values, collected by the Bureau of Labor Statistics;

4) Portfolio capital flows other than banking transactions, collected by the Treasury Department;

5) Banking transactions, collected by the Federal Reserve Board.

Several agencies, including the Customs Services, the Federal Reserve Board, its agent, the Federal Reserve Bank of New York, the actual collector of the banking data, and the Treasury Department are either unable to separate the costs of collecting current international data from the costs of regulatory activities or costs of collecting domestic data, or do not wish to divulge their expenditures. I am therefore unable to say much about the inputs of resources into these aspects of international data collection.

There are three questions I would like to answer about these data: First: what has happened to the input of resources into them? Second: what has happened to the amount and accuracy of information about each area? Third: are these data adequate to answer the questions asked of them, and has their adequacy been improving or declining?

I. Expenditures for International Data

The trends in resources into several elements of international data collection are described in Table 1, using expenditures deflated by the implicit deflator used by the Office of Management and Budget (OMB) for outlays other than defense and capital investment (1982 = 100). The general picture is of either stability of real resources or some increase, although the deflator I use is suspect, as I point out later.

TABLE 1[a]

Fiscal Years	Census Bureau Foreign Trade	BEA	BLS Int'l Prices
1977–81	13,624	4,277	3,966[b]
1982–85	12,470	4,768	3,997
1986–89	12,628	5,269	4,125
1990 Request	13,816	5,741	4,277

[a] $ thousands.
[b] 1981 only.

Input into the Census Bureau's foreign trade statistics program in 1986–89, as measured by appropriations, was about 7 percent below the 1977–81 level, and for 1982–89 was about 8 percent below. The estimate for fiscal 1990 in the president's budget message is at about the 1977–81 level.

The BEA international data collection effort, unlike the Census Bureau foreign trade statistics division, has seen gains in resources over the decade or so, by almost a quarter, most of which was for surveys. The BEA request for 1990 also represents a relatively large increase.

---

*Queens College and The Graduate Center, City University of New York, New York, NY 11367, and NBER. I am indebted for assistance with data and background material to Anne Y. Kester, Study Director of the Panel on Foreign Trade Statistics of the National Research Council, Brenda Gallagher of the U.S. Department of Commerce, and particularly to Paul Bugg of the Office of Management and Budget.

Real expenditures on the BLS international price program have crept up a bit since the late 1970s, but the growth has been slow. This program too would gain if the 1990 budget request ends up as appropriations.

In the case of the data on portfolio capital, the 1984 survey is reported to have cost much less even in nominal terms than the 1978 survey (U.S. Department of the Treasury, 1989, p. 8), and the difference is attributed to computerization. However, I have no information about the cost of the monthly Treasury International Capital reporting system.

II. The Quality of International Data

Despite the apparent decline or stagnation in resources, the quality of the trade data appears to have improved in some respects, particularly in the last few years. The so-called carry-over problem (the reporting of trade, particularly imports in the wrong month because the data processing fell behind) that sometimes reached 8–14 percent on exports, and as high as 50 percent or more on imports in late 1985 and early 1986, has been reduced very substantially. By the end of 1988, it was only 2 to 3 percent in unrevised data and under 1 percent in revised data (Bruce Walter, 1990, p. 7).

This improvement in quality in terms of timing was achieved partly by delaying the publication of the trade data by 2 weeks. For those of us who like to look at 100 years of data, the sacrifice was small, but there undoubtedly are some whose weighting of quality elements is different.

The Census Bureau has also recently announced, with some fanfare, a program for correcting the data on U.S. exports to Canada for omissions that amounted to $10 billion in 1986, mainly by using Canadian import data. I cannot resist pointing out that a very elaborate reconciliation of U.S. and Canadian data was published for 1970 with considerable commodity detail that has never been repeated. The U.S. authorities were aware of the underreporting of U.S. exports to Canada as early as 1867, and the study of the U.S. balance of payments by Matthew Simon (1960) estimated the omissions starting in 1861 by the ingenious method of using (guess what) Canadian import data. It took 120 years to solve this problem, which seems an unduly long time.

The expansion of resources in a particular area of data collection may not produce a gain in quality. The task to which the resources are directed may have become more complex. Export and import trade, for example, has roughly doubled in importance relative to GNP since the 1960s. The share of imports subject to tariffs has declined sharply, and will decline even more with the signing of the U.S.-Canada Free Trade Agreement. The system for correcting data on U.S. exports to Canada by using Canadian import data may not work as well once tariffs are removed, and some other method, requiring more resources, may be needed just to maintain the quality of the data. Similarly, the rising share of trade that is intrafirm may make present methods for calculating trade values (and trade prices also) less and less reliable, because there is a large arbitrary element in the prices and values reported.

Given that there has been some improvement in the quality of U.S. trade data in recent years, how was it achieved in the face of the combination of apparently constant or declining resources and increasing quantities of data to be handled? A major element was apparently the move toward computerization by customs brokers (presumably for their own purposes), by the Customs Service (not part of the statistical system), and by the Census Bureau itself (within the statistical system) to take advantage of the changes in the operations of the customs brokers and the Customs Service. This improvement in the quality of trade data implies that either total factor productivity was increasing in the statistical system, or the deflator I use gives a biased picture of resource inputs, or both. I am inclined to suspect that the latter has been most important. The deflator does not take account of the falling price of computing or of the rise in quality of the data supplied by other agencies or by respondents because they have been computerized.

There have been many improvements in the data on U.S. direct investment abroad, foreign direct investment in the United

States, and international service transactions. We now have annual reporting on the activities of U.S. direct investors abroad. In contrast to the 1966 Census of U.S. direct investment abroad (the final report that came out so late that no publication date is listed, and the preliminary reports that did not appear until 1971 and 1972), the preliminary report for 1987 and the final report for 1986 appeared in 1989. The current reports are now available on diskette, a clear gain for researchers.

Annual data on foreign direct investment in the United States have been published for the years since 1977 and that, too, represents a great improvement over what was available in the past. Overall, then, the increases in budget allocations to the BEA do appear to have led to improvements in the quantity and quality of data available on inward and outward direct investment.

Another area of clear improvement in international data is that of transactions in services, including the initiation of fairly comprehensive surveys of transactions with unaffiliated foreigners. Service transactions among U.S. firms and their affiliates were first segregated from goods transactions in 1982, and the same separation was extended to the U.S. affiliates of foreign firms in 1987 (Bernard Ascher and Obie Whichard, 1990, p. 36). Here, too, the statistical system seems to be gaining on its problems.

The BLS international price program has steadily increased in scope. No price data at all were available in the early 1960s, and only selected items were included in the first decade or so of the program. In the last 10 years, the program has come to provide fairly complete coverage for quarterly indexes of export and import prices.

In the case of portfolio capital stocks and flows, and the income resulting from them, the quality and accuracy of the data have almost certainly deteriorated over time. The deterioration has been a worldwide phenomenon that showed up in the enormous world net debit on current account that rose from \$10–20 billion in the 1970s to \$100 billion or so in the 1980s (Stephen Taylor, 1989, p. 401). The "Final Report..." of the IMF working party (1986) concluded that a large part of the problem involved income on portfolio investment and, by implication, the data on portfolio capital flows themselves. For the United States, in 1983, U.S. nonbank asset positions with foreign banks were \$53 billion in U.S. data and \$168 billion in reports by foreign banks to the IMF (Taylor, p. 416). These discrepancies are large enough to alter the impression of the U.S. current account, but they also imply that large flows of portfolio capital have taken place unnoticed by the U.S. statistical system. It is surely no help that the last benchmark survey of U.S. portfolio investment abroad was taken more than 45 years ago. This does seem to be one case in which the U.S. data on international transactions have deteriorated badly.

III. The Adequacy of International Transactions Data

In view of the improvements that have taken place, why is there so much dissatisfaction with the data on U.S. international transactions? While most of the data have improved in quality, the demands on them have increased faster than the improvement. The international transactions of the United States have become much more important over time to American policymakers and analysts. The rise in importance of international merchandise trade is one example of the increasing importance of international transactions, but, in addition, markets for services have been growing more rapidly than markets for goods, and the internationalization of capital markets has moved even faster than the growth of service industries in general. The better data of recent years (portfolio capital and portfolio income data being the exceptions) may still be less and less adequate for the demands put upon them.

The data on international service transactions, for example, were long known to be poorly estimated. The move by the United States to place services on the agenda for the Uruguay Round of GATT negotiations led to the need for better data to understand trade in services in order to form negotiating positions and to carry out negotiations. What had been a mere academic complaint became a deficiency felt by Congress and by administrative agencies, and it was this inadequacy for policy needs that led to the allocation of additional resources.

While Lois Steckler and Guy Stevens (1990) acknowledge the relatively high quality of U.S. data on direct investment transactions and operations, they point out many areas in which the data are nevertheless inadequate to answer important questions about international policy. The recent growth of foreign direct investment in the United States had produced greatly increased demands for information on how foreign companies operate here, and what the effects are on the American economy. A study of these questions would require the kind of detail on operations that could only be provided by a linking of the BEA data with those collected on an establishment basis by the Census Bureau in the *Census of Manufactures* and other business censuses. The current anxiety about the international investment position in the United States has made many observers dissatisfied with direct investment stock estimates based on book values (rather than on market values or current replacement cost), although the book values had been used for many years without much complaint other than the usual academic grumbling. In this case, the move toward more flexible exchange rates, and the large currency revaluations that followed, contributed to the discontent with book value data.

Thus, the relatively high grades we give the U.S. statistical system for improvements in the quality of international data must be reduced in many cases when we grade the system on the adequacy of the data for setting and analyzing of international economic policy. Where improvements have taken place, they seem to have stemmed from congressional pressures and, in some cases, from the consciences of the staffs of the agencies themselves. Successive administrations have shown little interest in the quality of the data they produce and presumably use for policymaking. What the system needs is some realization that the costs of wrong decisions based on misinformation are probably much larger than the costs of better information.

REFERENCES

Ascher, Bernard, and Whichard, Obie G., "Developing a Data System for International Sales of Services: Progress, Problems and Prospects," in P. Hooper and J. D. Richardson, eds., *International Economic Transactions*, forthcoming 1990.

Hooper, Peter and Richardson, J. David, *International Economic Transactions*, NBER *Studies in Income and Wealth*, Chicago: University of Chicago Press, forthcoming 1990.

Simon, Matthew, "The United States Balance of Payments, 1861–1900," in *Trends in the American Economy in the Nineteenth Century*, NBER *Studies in Income and Wealth*, Vol. 24, Princeton: Princeton University Press, 1960, 629–715.

Steckler, Lois and Stevens, Guy, "The Adequacy of U.S. Direct Investment Data," in P. Hooper and J. D. Richardson, eds., *International Economic Transactions*, forthcoming 1990.

Taylor, Stephen, "World Payments Imbalances and U.S. Statistics," in R. E. Lipsey and H. Stone Tice, eds., *The Measurement of Saving, Investment, and Wealth*, NBER *Studies in Income and Wealth*, Vol. 52, Chicago: University of Chicago Press, 1989.

Walter, Bruce C., "Quality Issues Affecting the Compilation of the U.S. Merchandise Trade Statistics," in P. Hooper and J. D. Richardson, eds., *International Economic Transactions*, forthcoming 1990.

International Monetary Fund, "Final Report of the Working Party on the Statistical Discrepancy in World Current Account Balances," Washington, December 1986.

U.S. Department of the Treasury, *Report on Foreign Portfolio Investment in the United States as of December 31, 1984*, Washington: Office of the Assistant Secretary, International Affairs, 1989.

Reviving the Federal Statistical System: A View From Within

By JACK E. TRIPLETT*

A number of recent reports, as well as articles in the press, have pointed to data gaps and inadequacies in the outputs of statistical agencies. Though one might argue some of the specifics, the content of these reports will evoke broad agreement among professional users of economic statistics. Budget cuts during the 1980s, alleged lack of "coordination and leadership" from the Office of Management and Budget (OMB), increasingly noncompetitive federal salaries for technical occupations, and increasing resistance by respondents to federal statistical surveys are listed as causes of the data gaps; curiously, criticism of statistical agencies themselves is muted.

Though it is undoubtedly true that statistical agencies could have made some headway toward filling data gaps had, for example, more resources been available in the 1980s, a longer view of the statistical system suggests that recent factors account for only a part of the data gap problem, and perhaps not the largest part. For one thing, complaints that statistical agencies do not anticipate and keep up with data needs are not new—they motivated, in part, the Bonnen Report 10 years ago, and were not new then. The federal statistical system has steadily lost (for good reasons and bad) its market share of the data used for economic research and for business purposes, and this decline is not uniquely a phenomenon of the 1980s (longitudinal data sets, for example, were new products developed largely outside statistical agencies, partly because of agency antipathy).

Second, many of the unmet data needs documented in recent reports have long been apparent. For example, all agree that service sector data are inadequate; however, this is hardly a new lacuna in statistical agency outputs. The proportion of the U.S. labor force engaged in the non-goods-producing industries surpassed 50 percent in 1940; and over 20 years ago, Victor Fuchs emphasized both the increasing importance of the "service economy," and its lagging data development. Yet, little service sector data (and minimal planning for it) existed by 1980. Factors affecting statistical agencies during the 1980s contribute little to explaining, in the sense of scientific causality, why U.S. data for services remain inadequate in 1990. Similar statements, unfortunately, can be made about other gaps.

I focus on the "long term" question that is relatively neglected in recent reports: Why does the statistical system respond so slowly to developing needs, and what might be done about it? Reversing the usual order, I present the prescription first, followed by the diagnosis.

I. What Might Be Done?

1) Statistical agencies need increased capability for research and analysis. Though the most important activity in a statistical agency is to get out the monthly or quarterly numbers and to get them right, research on concepts and methods (mentioned in many reports on the statistical system) serves design and development functions in statistical agencies. Moreover, because research points to emerging needs for economic data, the role of analysis in statistical agencies must be expanded to use researchers' skills in the program planning process. Increasing the role of research, analysis, and evaluation is the most important single contribution toward making statistical agencies more responsive to emerging needs.

---

*Chief Economist, U.S. Bureau of Economic Analysis, Washington, D.C. 20230. Views are my own, do not represent official positions of either the Bureau of Economic Analysis or the Department of Commerce, and are not necessarily those of other staff members of those agencies. A longer version of the paper, containing references and footnotes plus additional development of the proposals, is available on request.

2) Mechanisms for the integration of disparate data sets produced by different statistical agencies need to be improved; increased links are needed between collecting and compiling agencies and governmental integrating units, including policy analysis units outside the statistical system. Data integration is inherently an analytic activity. It promotes a broad view of data needs and focuses attention on data gaps and inconsistencies among series. Locating budgetary authority for incremental data needed for integrated programs with the integrating agency, rather than solely with the collecting agencies (the traditional practice), would be a useful step. Despite some inherent dangers, I favor the "purchase" of statistical data by nonstatistical government analytic units that are major users. This proposal is intended to increase contact between statistical agencies and analytic users of their data, thereby making program managers more aware of analytic needs; it also reduces the amount of coordination that is required from the statistical arm of OMB, because it circumvents some (but not all) of the worst managerial problems that arise from the decentralization of the U.S. statistical system.

3) Professional advisory groups are important to statistical agencies. Each statistical agency should have one. Advisory committees should be composed of analytic data users, with representation as well for respondents to government surveys. I recommend further that the administrative responsibility for advisory committees be lodged in statistical agency analytic units, partly for the sake of implementing the advisory committee's recommendations, and partly to enhance the "clout" of analytic staffs within the agency. Moreover, the advisory group should explicitly be charged with identifying low-priority statistical programs whose elimination would improve the relevance of statistical agency outputs. This proposal does not preclude the formation of special outside review groups to examine particular programs or issues of statistical methodology.

4) Gerard Debreu has asked: How can professional associations help improve the performance of the statistical system? I urge professional societies not only to encourage more measurement research (a recommendation that appears in other reports on the statistical system), but also to take steps to improve the quality of such research, with much more attention to replication and verification. Outreach activities to assure that statistical agency professionals are not inappropriately isolated from researchers in other employments would be valuable. (As a corollary, societies should bestow professional honors on statistical agency staff only for professional accomplishments, and not merely because they hold high-level administrative jobs.)

The content of the first two proposals parallels major recommendations of the Bonnen Report. I propose, however, locating these activities for the most part within statistical agencies themselves, or in major user groups, rather than in a greatly enlarged statistical unit in OMB. One reason is my belief that it is desirable to locate research as close to data production as practicable; in addition, I do not regard a large OMB analytic unit for *statistical* purposes as likely in the foreseeable future, even if OMB were a preferable location for these activities. The longer version of this paper contains a more complete statement of the proposals and their rationale.

II. Why Does the System Respond So Slowly?

The justification for the federal government's production of economic statistics eventually turns on economic policy needs. Recently, Charles Schultze asserted that the specification of data needs for economic policy must originate from economic and social research agendas, because research provides the information that is ultimately decisive for policy making. Though the federal government should not necessarily supply all research data, the key role Schultze has assigned to research is a view with which I am wholly in agreement.

Indeed, I would go further. Today's research need for data becomes tomorrow's policy-analytic need (see as an example Robert Lipsey's "academic grumblings" remark in his paper for this session); and today's research methodology (which implies a specification for the way data should be organized) is the method employed for tomorrow's policy analysis. Moreover, the technical gap between what is done in the more sophisticated Washington policy analysis shops and the state-of-the-art in academic research has greatly narrowed in recent years.

Yet, with a few distinguished exceptions, statistical agency program managers and administrators are unfamiliar with economic policy analysis, and have neither the inclination nor the ability to fathom data needs for basic economic research. An unfortunately pertinent example is an administrator who adamantly insisted in the mid-1980s that the demand for longitudinal data had "peaked," and determined that it should receive *no* priority in labor market data programs.

Given the advantages from specialization of labor, it is neither necessary nor desirable that statistical program managers all be economic researchers. Nevertheless, the first problem is: The voice of analytic and research needs for data is very weak in statistical agencies.

The second causal factor is a variation on the first: The statistical system's information on needs is filtered through aspects of the Washington policy-making environment. The policy-making environment has had a strong bias toward macroeconomic forecasting of quarterly or monthly numbers. As a result, a statistical program classified as a "principal macroeconomic indicator" has implicitly been regarded as more deserving of budgetary support than one that is not. In the mid-1970s, for example, the relatively noncyclical nature of many service industries was one reason for rejecting, within one agency, a proposal for service sector data (thus, went the argument, service sector data made no contribution to macroeconomic indicators). Obviously, a priority-setting system that relies on an existing set of macro indicators is one with a great deal of inertia in it, and is biased as well. Macro data needs are important, but they are not the only needs.

In the Washington policy environment, many policy issues represent doctrinaire disputes on outdated programs. One can find, for example, economists who make persuasive cases that U.S. farm programs do not raise returns to farming, that middle-income homeowners' mortgage programs do not lower housing costs to middle-income homeowners, and that government labor market and training programs do not increase worker incomes. Yet, the policy debate usually concerns not whether 20, 40, or 60 years of program failure is evidence itself, but whether, for example, the labor force "needs" more training, with an affirmative answer taken as justifying an expanded federal government role. For the statistical system, this means, for example, resources put into labor market projections and data programs that supposedly help program administrators "correct" deficiencies in the operation of labor markets. Statistical agencies cannot resist such programs entirely, for the legislative process is the way the public directs the federal government; but agencies too frequently embrace them. Such marginal or counterproductive programs may soldier on for decades, sapping budgetary resources that could be used more productively, and more importantly, adding to respondent burden, thus provoking resistance to new collection.

Another aspect of the policy environment is represented by periodic hullaballoos over naive issues (such as the recent debate over whether the United States is "losing the good jobs," and the demand 10 years ago for measures of "labor market related hardship"). Such issues often do not turn into legislation, but they frequently result in demands on statistical agencies—often for some statistic that could be enlisted to support a particular policy agenda. Healthy debate is the way ideas are evaluated in a democracy. Statistical agencies, however, need the intellectual insight to see beyond the issues that thrive in the press—and that can be especially difficult when (as was true of both of my examples) issues that deserve to be ignored originate from respected academic research institutions!

In summary, statistical agencies' perceptions of the demands for policy-related data are dominated by macro data and by current issues that garner space in the newspapers. Some of these are far removed from, and antithetical to, long-run needs for research and analysis, and thus (if one accepts Schultze's characterization of the central role of research on policy making) are ultimately antithetical to the development of the data needed for economic policy making.

Many reviews of the U.S. statistical system point to the alleged lack of coordination or "leadership" from OMB, rather than factors within statistical agencies, as a reason that the statistical system has failed to anticipate needs for economic policy analysis.

There is something to the coordination complaint. Consider consumption statistics. The Bureau of Labor Statistics' *Consumer Expenditure Survey*, which contains both consumption and demographic information, provides a general consumption data base for policy analysis (tax-policy modeling, for example). A companion *Point of Purchase* survey collects information on the kinds of stores where households purchase, for example, durables. The federal government also conducts many special-purpose consumption surveys (such as household energy consumption surveys). The surveys are not drawn from common sampling frames by methods that would permit integration; sample totals are "blown up" to economywide estimates by counts of consumer units (rather than by controlled totals of consumption, such as GNP Personal Consumption Expenditures, less its nonhousehold component); and because the BLS regards its surveys primarily as inputs to its own Consumer Price Index, none of the potentially useful data from the *Point of Purchase* survey has ever been published.

Government consumption data exemplify "statistical coordination" problems—each decentralized collecting agency pursues its own program needs, without sufficient attention to integrating the data for analysis. But on another level, consumption statistics problems reflect lack of analytic input—they reveal a producer perspective, when a user one is needed. An energetic OMB unit with a user perspective would undoubtedly act; but so would statistical agencies, individually and in concert, had they more analytic perspectives in their own statistical planning processes. Integration of government consumption data was raised, after all, by Houthakker and Taylor 20 years ago.

Though it is undoubtedly true that an analytic unit within OMB might have some positive effect on the statistical system as a whole, statistical agencies need their own analytic capabilities in any case. Hence, my diagnosis: Statistical agencies have too few of the professionals who can provide insight into analytical needs, and when agencies have such professionals, they pay insufficient attention to them in the planning process.

The Development of Industrial Research in U.S. Manufacturing

By DAVID C. MOWERY*

Historical analysis of the development of industrial research in the U.S. economy raises important issues for economic historians, theorists, and scholars of industrial organization. This analysis suggests the importance of the utilization and transfer of *R&D* results, an issue that is largely overlooked in neoclassical economic theory's emphasis on the appropriability of the results of research (W. M. Cohen and D. Levinthal, 1989, is an exception). Economic theory also devotes little attention to the organization of the innovation process. Yet, the organization of research has influenced the performance of national systems of innovation and ultimately of national economies.

I. The Origins of Industrial Research

The expansion of the American economy during the late nineteenth and early twentieth centuries combined with innovations in transportation, communications, and production technologies to yield manufacturing operations of unprecedented scale (A. D. Chandler, 1977). The materials analysis and quality control laboratories that were established within many of these factories were among the first industrial employers of research scientists and engineers (N. Rosenberg, 1985; W. D. Lewis, 1967). These plant-level laboratories were gradually supplemented by laboratories devoted to longer-term research. Although many early quality control laboratories were responses to changes in the scale of production, the development of corporate research laboratories reflected change in the organizational structure of the firm. The development of industrial research was associated with expansion and diversification in the activities and products of manufacturing firms, and with the substitution of intrafirm for market control of these activities.

Structural change in many large U.S. manufacturing firms and their investment in industrial research were influenced by U.S. antitrust policy. The increasingly stringent judicial interpretation of the Sherman Antitrust Act in the late nineteenth century made agreements among competing firms for the control of prices and output more frequent targets of civil prosecution. The U.S. merger wave of 1895–1904, particularly the surge in mergers after 1898, was in part a response to this new legal environment (G. J. Stigler, 1968). Many of the firms created through these mergers rationalized their internal structure and developed a central office to coordinate the activities of their many components. In firms such as AT&T, General Electric, or Du Pont, the strengthening of a central office was associated with the establishment or significant expansion of a central research facility.

In other industrial economies at this time, such as Great Britain, weak antitrust policies reduced firms' incentives to merge and prevented the rationalization of the internal structure of the firms emerging from such mergers as did occur. One result was levels of *R&D* employment and investment in large

---

[†] *Discussants*: Paul Hohenberg, Rensselaer Polytechnic Institute; J. David Roessner, Georgia Institute of Technology; Melvin Kranzberg, Georgia Institute of Technology.

*Walter A. Haas School of Business, University of California, Berkeley, CA 94720. Support for this research was provided by the Division of Policy Research and Analysis of the NSF, and by the Technology and Economic Growth Program of Stanford University's Center for Economic Policy Research. I am indebted to Richard Nelson, Nathan Rosenberg, and Keith Pavitt for conversations that have contributed to many of the ideas in this paper.

British firms throughout the 1900–50 period that were below those of U.S. firms (see my 1984 paper). Nevertheless, a weak antitrust climate in other nations, such as Germany, was associated with growth in industrial research, making it difficult to assert a direct cause-and-effect relationship between antitrust policy and the growth of intrafirm *R&D*.[1]

The influence of antitrust policy on the growth of industrial research, however, extended beyond its effects on corporate structure. The incentives created by the Sherman Act for corporate mergers were reduced significantly after the *Northern Securities* decision of 1904. Judicial interpretation of the Sherman Act and Justice Department prosecution of a widening array of firms nevertheless led to increased reliance on industrial research and innovation to forestall or offset the effects of antitrust prosecution.

Firms threatened by prosecution under the Sherman Act used industrial research to diversify out of their primary industry, and to accumulate patents that could sustain or protect a dominant market position. Du Pont's research activities focused increasingly on diversification out of the black and smokeless powder businesses, even before the antitrust decision of 1913 that forced the divestiture of a portion of the firm's operations (D. Hounshell and J. K. Smith, 1989). Research laboratories were not concerned solely with internal inventive activity; they also monitored the environment for technological threats and opportunities for the acquisition of patents or firms (W. F. Mueller, 1962; Hounshell-Smith). The research laboratories of AT&T, General Electric, and Eastman Kodak served a similar monitoring function during this period.

Monitoring of the technological environment was facilitated, and the incentives for the establishment of industrial research laboratories were increased, by strengthened protection for intellectual property in the late nineteenth century. Federal court decisions in the 1890s upholding the validity of patents covering goods not in production increased the utility of patent portfolios for defensive purposes. The effects of these judicial decisions were reinforced by federal courts' acceptance of patent licensing in antitrust suits during the early 1900s. The 1911 consent decree settling the federal government's antitrust suit against General Electric left GE's patent licensing scheme untouched, enabling the firm to maintain an effective cartel within the U.S. electric lamp market for years to come without having to acquire competitors, an action prohibited by the consent decree.

Antitrust litigation in the early twentieth century influenced the research activities of other large U.S. firms. AT&T's 1913 commitment to the Justice Department by vice president Kingsbury foreclosed expansion through acquisition of other telephone operating companies, and meant that future growth would depend on innovation in telephone services and equipment, or diversification into related lines of business (L. S. Reich, 1985). The firm broadened its research activities, increasing the emphasis placed on internal innovations, and expanded its research activities in radio.

J. A. Schumpeter argued (1954) that organized industrial research had supplanted the inventor-entrepreneur (a hypothesis supported by J. Schmookler, 1957) and would reinforce the position of dominant firms. During 1921–46 at least, the effects of industrial research on turnover among the 200 largest U.S. manufacturing firms were consistent with his predictions. Industrial research significantly aided U.S. manufacturing firms' survival in the ranks of the 200 largest firms during this period (see my 1983a paper). The growth of industrial research during 1921–46 among the 200 largest firms was associated with a decline in turnover within this group. To the extent that federal antitrust policy contributed to industrial research investment by large, oligopolistic firms during this period, the policy paradoxically may have aided the survival of the largest firms.

---

[1] N. Lamoreaux (1985) argues that other factors, including the increasing capital intensity of production technologies and the resulting rise in fixed costs, were more important influences on the U.S. merger wave than antitrust policy, but her account (p. 109) acknowledges the importance of the Sherman Act in the peak of the merger wave. Lamoreaux also emphasizes the incentives created by tighter Sherman Act enforcement after 1904 for firms to pursue alternatives to merger or cartelization as strategies for attaining or preserving market power (p. 181).

II. The Role of Intrafirm and Contract Research

Industrial research in the United States developed a dualistic structure during its formative years. Employment of professional scientists and engineers grew rapidly within manufacturing concerns, and in independent research firms not affiliated with a manufacturing enterprise. Between 1900 and 1940, nearly 350 independent laboratories were established (see my 1983b paper). Independent research laboratories employed more than 5,000 scientists and engineers by 1946, although their share of total employment of research scientists and engineers declined during the 1921–46 period.

Although viewed by contemporary observers as a means for small firms lacking internal *R&D* laboratories to exploit industrial research (J. J. Carty, 1916) and by scholars as a logical extension of the division of labor (Stigler, 1956), independent research organizations do not appear to have substituted for in-house research during this period (see my 1983b paper). These organizations did not offer a full menu of research services for purchase, and were especially limited in product *R&D*. Limitations on the role of independent laboratories reflected problems of contracting, the importance of firm-specific knowledge that often is not codified, and difficulty in transferring and absorbing knowledge across organizational boundaries (O. E. Williamson, 1975, 1985; D. J. Teece, 1982).

As in-house research facilities grew in size and number during 1900–40, contract research institutions increasingly complemented, rather than substituted for, in-house industrial research. Firms lacking in-house research capabilities thus faced significant handicaps in the innovation process.

III. Higher Education and Industrial Research, 1900–40

To a greater extent than was true of Germany or Great Britain, industrial and academic research developed in parallel in the United States. The pursuit of research was recognized as an important professional activity within both the U.S. industry and higher education only in the late nineteenth century, and research in both venues was influenced by the example (and in the case of U.S. industry, by the competitive pressure) of German industry and academia.

Linkages between academic and industrial research were powerfully influenced by the decentralized nature of the structure and funding of U.S. higher education, especially public universities. Public funding helped support a higher education system in the United States that was far larger than that of such European nations as Great Britain. Of equal or greater importance, however, was the fact that public funding for many U.S. universities was provided by state governments, rather than by the federal government. Indeed, the contribution of state governments to nonagricultural university research appears to have exceeded the federal contribution during the 1930s (see my study with Rosenberg, 1989).

Both the curriculum and research activities of U.S. public universities were more closely geared to commercial opportunities than was true in many European systems of higher education. Especially within emerging subfields of engineering, state university systems often introduced new programs as soon as the requirements of the local economy became clear. The first state-supported engineering experiment stations were established early in the twentieth century, and numbered 38 by 1938. Much of the work conducted by the engineering extension stations was applied in character, rather than focusing on basic research.[2]

IV. Conclusion

This historical overview of the development of U.S. industrial research indicates that the structure of the institutions performing *R&D* matters a great deal. The characteristics of the learning processes that underpin innovation, and the knowledge produced by research, led to the development of industrial research within the United States and other industrial economies as an in-house activity, rather than as a commodity sold through arms-length contracts in a market. Where institutional and other forces discouraged the growth of research within the firm, investment in *R&D* was lower. The complementarity of in-house and external research is consistent with the activities of many in-house industrial research laboratories in technology monitoring and acquisition.

---

[2] I am indebted to Nathan Rosenberg for this point.

The structure of the twentieth-century U.S. *R&D* system of which the industrial research laboratory was a central component now may be changing. Many U.S. firms are exploring alternatives to exclusive reliance on intrafirm sources of innovation. The discussion above suggests that these experiments are not entirely novel. Indeed, the recent upsurge in state government funding of applied research and the proliferation of collaborative research relationships between universities and industry represent a partial revival of earlier relationships that were sundered by the dramatic changes in the structure of the U.S. research system during and after World War II.

Rosenberg (1982) argued that the industrial research laboratory made the scientific research agenda more responsive to economic factors and therefore "endogenized" science to some degree. Much of the current experimentation with new organizational forms for industrial research, however, may be responding to a very different trend. In some sectors, the in-house industrial research laboratory can no longer influence the direction of scientific research and thereby exploit linkages between basic and applied research as effectively as before. Instead, many U.S. firms now are trying to develop links to external *R&D* as a means of searching their technological environment more aggressively, in order to adapt to the appearance of new and commercially relevant scientific and technological advances from external sources.

REFERENCES

Carty, J. J., "The Relation of Pure Science to Industrial Research," *Science*, October 13, 1916, *44*, 511–17.

Chandler, A. D., Jr., *The Visible Hand*, Cambridge: Harvard University Press, 1977.

Cohen, W. M. and Levinthal, D., "Innovation and Learning: The Two Faces of R&D," *Economic Journal*, September 1989, *99*, 569–96.

Hounshell, D. and Smith, J. K., *Science and Corporate Strategy*, New York: Cambridge University Press, 1989.

Lamoreaux, N., *The Great Merger Movement in American Business, 1895–1904*, Princeton: Princeton University Press, 1985.

Lewis, W. D., "Industrial Research and Development," in M. Kranzberg and C. Pursell, eds., *Technology in Western Civilization*, New York: Oxford University Press, 1967.

Mowery, D. C., (1983a) "Industrial Research and Firm Size, Survival, and Growth in American Manufacturing, 1921–46: An Assessment," *Journal of Economic History*, December 1983, *43*, 953–80.

______, (1983b) "The Relationship between the Contractual and Intrafirm Forms of Industrial Research in American Manufacturing, 1900–1940," *Explorations in Economic History*, October 1983, *20*, 351–74.

______, "Firm Structure, Government Policy, and the Organization of Industrial Research: Great Britain and the United States, 1900–1950," *Business History Review*, Winter 1984, *53*, 504–31.

______ **and Rosenberg, N.,** *Technology and the Pursuit of Economic Growth*, New York: Cambridge University Press, 1989.

Mueller, W. F., "The Origins of the Basic Inventions Underlying Du Pont's Major Product and Process Innovations, 1920 to 1950," in *The Rate and Direction of Inventive Activity*, Princeton: Princeton University Press, 1962.

Reich, L. S., *The Making of American Industrial Research*, New York: Cambridge University Press, 1985.

Rosenberg, N., "How Exogenous is Science?," in his *Inside the Black Box*, New York: Cambridge University Press, 1982.

______, "The Commercial Exploitation of Science by American Industry," in K. B. Clark et al., eds., *The Uneasy Alliance*, Boston: Harvard Business School, 1985.

Schmookler, J., "Inventors Past and Present," *Review of Economics and Statistics*, August 1957, *39*, 321–33.

Schumpeter, J. A., *Capitalism, Socialism and Democracy*, 3d. ed., New York: Harper & Row, 1954.

Stigler, G. J., "Industrial Organization and Economic Progress," in L. D. White, ed., *The State of the Social Sciences*, Chicago: University of Chicago Press, 1956.

______, "Monopoly and Oligopoly by Merger," in his *The Organization of Industry*, Homewood: Irwin, 1968.

Teece, D. J., "Towards an Economic Theory of the Multiproduct Firm," *Journal of Economic Behavior and Organization*, March 1982, *3*, 39–63.

Williamson, O. E., *Markets and Hierarchies*, New York: Free Press, 1975.

______, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

Punctuated Equilibria and Technological Progress

By JOEL MOKYR*

Charles Darwin and Alfred Marshall were both extremely influential men, whose influence has waned somewhat in recent decades without diminishing in any sense their stature in their respective disciplines. Marshall and Darwin shared a mutual interest in each other's science. Marshall suggested that "the Mecca of the economist lies in economic biology," though he never explained what he exactly meant by that. Recently, economists and biologists have found some common ground, as in their shared interest in game theory (John Maynard Smith, 1982), and in adaptive system analysis of disequilibria (Richard Day, 1987). Analogies between economics and biological sciences have become more popular among economists in recent years, and the number of books and journals published with the word "evolution" figuring in the title is rising rapidly. Physical and mechanical metaphors are coming under fire (Philip Mirowski, 1990).

One isomorphism that thus far seems to have escaped notice is that both economic historians and evolutionary biologists have been debating for years whether past changes in the objects of their studies have been smooth and gradual or whether history has moved in leaps and bounds. Darwin and Marshall both believed that nature does not make leaps. Both were influenced by a long and venerable tradition that harked back to Leibniz rooted in the Aristotelian notion of the continuity of space and time (Alexander Gerschenkron, 1968). Darwin and Marshall, however, were concerned with information systems, and here the absence of leaps is less obvious.

The term evolution, after all, has two distinct meanings: one is a dynamic system of natural selection imposed on blind variation; the other is evolution as distinct from revolution, gradual as opposed to abrupt. But how evolutionary was evolution? Economic historians, and their equivalents in the biological world, the paleontologists and population geneticists, have been debating for a long time whether the past demonstrates that change is gradual or not, and whether at times novelties emerged without clear-cut precedence. Recent thinking among historians interested in the history of technology and long-term economic change seems to favor once again the basic notions of gradualness (Eric Jones, 1988) and precedence (George Basalla, 1988). In what follows, I reexamine the "continuity in history" question in the context of technological change, using the analogy with evolutionary biology as a mirror.

What genetics and technology have in common is, above all, that they are decentralized, self-organized systems of information that are transmitted from generation to generation. The differences between them are quite obvious: biological information is transmitted through parental genes, and acquired characteristics are not passed on. In all cultural evolution, including technology, information is transmitted laterally and diagonally as well as vertically. Cross-lineage hybridization, common in all cultural evolution systems, is rare in the biological world. The time scale of change is totally different. There are complex and subtle difficulties with the analogy, some of which I discuss elsewhere (1989, 1990b). All the same, the systems are inherently comparable and something can be learned from the analogy.

Information systems resist change, or else they would quickly degenerate into anarchy.[1]

---

*Departments of Economics and History, Northwestern University, Evanston, IL 60208. The comments of Louis Cain, Paul David, and Steven Matthews are gratefully acknowledged.

[1] In biological change, the vast majority of all mutations are weeded out by nature as birth defects. Moreover, structural constraints limit the extent to which new genotypes can differ from the status quo. In techno-

Yet total stability is not a defining characteristic, or else there would be no history. Observed changes in genetic and technological systems can be classified into four different classes:

a) Phenotypical changes without genotypical causes. Much observed variation is not the result of different information, but of a reversible response of the phenotype (visible features) to changes in environment, such as shedding a fur. The equivalent in economics is a movement along a production function, that is, a choice between known technologies responding for instance to changes in factor prices or to seasonality.

b) Changes in gene frequency or dispersion of existing information, a process governed by natural selection. Given that there is some genetic variation to start with, the genes of the species with the highest fitness (i.e., reproducing the fastest) will come to dominate the population. This will be recognized readily as the analogue of the diffusion process of new technologies. Diffusion will eventually peter out if no new inventions are forthcoming. The description of natural selection by a population geneticist as "a fire that consumes its own fuel" (Richard Lewontin, 1982, p. 151) is thus apt for both processes.

c) Mutation, a change in the genotype, is analogous to the emergence of new ideas. Here, too, the analogy should not be interpreted literally because in nature mutations occur as copying errors in the DNA, whereas new ideas, though highly stochastic, have an element of intentional directionality in them. Most mutations are rather small and can hardly be regarded as the beginning of new species, just as most inventions are individually insignificant cumulative improvements on existing technologies.

d) A very small minority of mutations result in the generation of new species, a process known as speciation. A felicitous metaphor coined by the geneticist Richard Goldschmidt (1940) for macromutations (that lead to the evolvement of a new species) is "hopeful monsters." The metaphor seems as apt a description of an early model of the Newcomen steam engine or the early power loom as one of a mutant living being. Hopeful monsters are rare in technology as they are in living species, and successful ones far rarer. By analogy, I will define macroinventions as technological breakthroughs that constitute discontinuous leaps in the information set and create new techniques.

Changes of types (a), (b), and (c) are consistent with the gradualist axioms. Phenotypical change tends to be short-run in nature, and rarely involves very dramatic changes. Switching to a different kind of grain in response to a pest, or to a more labor-intensive technique as a response of population growth, are examples of such changes. Similarly, the continuity of diffusion processes is preserved in our S-shaped diffusion curves.

The emergence of new techniques, however, has at times taken the form of dramatic changes. Such heroic interpretations of history have been properly criticized as oversimplified, but they contain a kernel of historical truth. The biological analogy of such abrupt changes is speciation. Speciation played a relatively small role in Darwin's biology, but recent developments have restored it to its rightful place in the center of evolution. Paleontologists such as Gould, Eldredge, and Stanley have rebelled against the ruling neo-Darwinian orthodoxy that insisted that the evolutionary process works in small and continuous increments. Instead, they maintain, Goldschmidt was fundamentally right when he argued that evolution proceeded in fits and starts (Stephen Gould, 1982). Long periods of stagnation and very slow gradual change were punctuated by feverishly rapid changes in which large numbers of new species emerged, a process known as adaptive radiation.

My argument is simply that much of the record of the economic history of technology displays a similar dynamic pattern of long

---

logical systems, the equivalent of structural genetic constraints are production externalities recently emphasized by Paul David (1988) and Brian Arthur (1989). Furthermore, as technological progress is never Pareto superior, resistance to change often resulted in Luddism, and political struggles between winners and losers (Mancur Olson, 1982).

periods of stagnation or very slow change, punctuated by sudden outbursts like the Industrial Revolution. Evolutionary systems driven by the interaction of radical innovations and smaller innovations, and diffused by natural selection do not necessarily advance in a smooth and gradual way. Self-organized information systems tend to be nonlinear, and cannot be characterized in any sense by linear stability.

A few remarks are in order about these macroinventions. Some of them were genuine "eureka" type of breakthroughs in which a single inventor revolutionized a whole technical mode in one bold stroke. In other cases, a second invention of major importance had to be made to perfect the first macroinvention.[2] The relation between macro- and microinventions is one of complementarity. Such complementarities occur in biological evolution as well, although they take a different form. It seems incomplete, however, to argue only that "the major innovations are made possible by numerous minor innovations" (Devendra Sahal, 1981, p. 37). Instead, the reverse seems more important. Radical advances in the manipulation or understanding of physical processes are usually the beginning, not the end, of a prolonged process of improvements and modifications.

While these microinventions may well have been the principal source of measured productivity growth, without the original macroinvention there would be nothing to improve. The chief importance of radical inventions is that they raise the marginal product of effort in development, and thus lead to a sequence of further improvements. An implication of this theory is that, in periods of radical inventions, we observe an intensification of smaller inventions as well.

On the other hand, a radical insight is not enough. Just as a mutant who survives but cannot reproduce is not a successful mutation, the many inventions made by Leonardo, and seventeenth-century curiosa such as Stevin's submarine, Branca's steam turbine, and Pascal's calculator cannot be regarded as successful macroinventions because they could not be produced at acceptable costs. Lewis and Paul's 1740 invention of the roller replacing human fingers as a yarn-twisting device had to wait until it was complemented a quarter-century later by Arkwright's relatively marginal but crucial insight to use two (instead of one) sets of rollers. Bessemer's invention of the converter would have been useless had it not been for Robert Mushet's addition of *spiegeleisen* (an alloy of carbon, manganese, and iron) into the molten iron as a recarburizer. Such information complementarities are parallel to the complementarities implicit in biparental reproduction and the phenotypic expression of mutations in recessive genes.

How does one distinguish between macro- and microinventions? Devising a metric that maps a single value of "radicalness" of the novel information from the set of complex characteristics of each invention is, of course, impossible. Although no such exact metric can be devised, some examples should help.

The later middle ages in the West witnessed a number of macroinventions: the weight-driven clock, the windmill, and the blast furnace. These three inventions created techniques that differed enormously from previous usages. The weight-driven clock used a verge and foliot escapement to convert time, a continuous and unidirectional variable, into oscillating and bidirectional motion. The windmill was the first use of wind power for anything except sailing and represented a revolution in power transmission. The blast furnace completely revolutionized the iron smelting process and made its casting possible.

Most macroinventions between 1500 and 1750 were stillborn, in part because of structural constraints on workmanship and materials. By 1750, the environment changed and the Industrial Revolution was, in its barest essence, a clustering of macroinventions (see my 1990a paper). By the criterion of the novelty of information, the quintessential inventions of the Industrial Revolution were not the mule or the rolling-and-puddling process, but hot air ballooning and gaslighting. The clustering phenomenon of radical innovations is widely observed in all cultural processes, and represents a combination of a conducive environment and interactions between the agents themselves. In biology, the agents do not respond to mutations in other species directly, yet each species takes changes in others as changes in its environment. These interactions, too, may result in critical mass situations leading to occasionally intensive and sudden outbreaks of speciation.

---

[2] Thus the addition of a separate condenser to an atmospheric engine or the four-stroke principle to the internal combustion engine must be counted as part of the macroinvention. The weight-driven clock was vastly improved by the pendulum.

The best examples for macroinventions come, understandably, from the second Industrial Revolution of the later nineteenth-century. One example that neatly illustrates the differences between micro- and macroinventions was telegraphy. Before the 1830s, information could not travel faster than people with the exception of homing pigeons and semaphore, both clumsy and unreliable systems. Telegraphy, like many other macroinventions, cannot be attributed to a single inventor, even though it evolved in a short period. The idea that information could travel at the amazing speed of an electric current occurred in the first decades of the nineteenth century.[3] Decades of gradual improvements followed.

The first commercial telegraph line was established in London by William Cooke and Charles Wheatstone in 1837, and the first successful submarine cable between Dover and Calais was laid in 1851. Yet, long-distance cables were found to be a difficult technology to master. Signals were often weak and slow, and the messages distorted. Distortion increased with distance, a problem known as capacitance. Moreover, submarine cables were subject at first to intolerable wear and tear. Physicists, above all William Thomson (later Lord Kelvin), made fundamental contributions to the technology.[4] By 1910, cable technology was still imperfect. In the 1910s, automatic relays were installed, and later complemented by repeaters that sampled the incoming signal, greatly increasing accuracy and speed. At about the same time, it was recognized that by wrapping the core of cables in an alloy of soft iron and nickel (a process known as "loading"), cables could overcome the problem of capacitance and carry signals further and three to five times faster than before. Compared to the breakthroughs in the 1830s, however, these inventions were marginal.

The last stages of the continuous improvement in cable technology was more or less contemporaneous with the development of the wireless by Guglielmo Marconi. Marconi's success in using Hertzian waves for the transmission of information and applying Oliver Lodge's idea of syntony to fine-tune transmissions has been well-described (Hugh Aitken, 1976). In his work in the 1890s, Marconi worked with long-wave radio, as it was found that long waves followed the contours of land. For long distance communications, long-wave radio needed enormously large and expensive antennas, and could not compete with cable.

Short waves, because their behavior was more like light waves, were long believed to be useless in long-distance communications. Amateurs kept, however, picking up signals transmitted from inexplicable long distances. In February 1924, Marconi announced a startling finding: shortwaves were bounced off the ionosphere, and thus could be used for long-distance telegraphy. Marconi thus "changed the world twice" as Headrick has remarked (1990). Shortwave was to radio what the separate condenser was to steam, the second stage of a radical breakthrough. The telegraph as a new specie was born in the 1830s; the birth of long-wave radio in the 1900s and of shortwave radio in the 1920s were equally revolutionary.

---

[3] In 1820, André Ampère suggested using the deviation of a magnetic needle in a telegraph, and in the same year François Arago established the principle of the electromagnet. Together, these two insights formed the theoretical basis of telegraphy.

[4] In the 1850s, Thomson invented a special galvanometer, and a technique of sending short reverse pulses immediately following the main pulse, to sharpen the signal (Daniel Headrick, 1990, pp. 215–18).

Long-distance communications thus illustrate the abruptness of technological change. There was no natural transition from semaphore to the electrical telegraph, nor a gradual movement from the telegraph to the first radio transmission by Marconi in 1894, nor a smooth natural development from the long-wave radio used in the first thirty years to the shortwave systems of the later 1920s. Each of the three systems was subsequently perfected by a long sequence of microinventions, but these would not have occurred without the initial breakthroughs. Their concept was novel, they made things possible that were previously impossible, and they were pregnant of more to come. Therein lies the essence of a macroinvention.

We know relatively little about the causes of macroinventions. Here, too, the similarity with evolutionary biology holds up. Whereas microinventions largely result from the responses of research and development to market forces and opportunities, the original events creating those opportunities are less readily explained. Many macroinventions, just like the emergence of species, were the result of chance discoveries, luck, and inspiration. Biologists agree that certain environments are more conducive to speciation than others. For example, speciation tends to occur in relatively small and isolated populations. Similarly, we can identify (though with a much greater margin of uncertainty) elements that contribute to an environment conducive to macroinventions. Such a search for the true causes of change remains speculative, in both the economic history of technology and evolutionary biology. We cannot hope, however, to understand the historical changes that really mattered without realizing that nature makes leaps, from time to time.

REFERENCES

Aitken, Hugh G. J., *Syntony and Spark: The Origins of Radio*, New York: Wiley & Sons, 1976.

Arthur, Brian, "Competing Technologies, Increasing Returns, and Lock-in by Historical Events," *Economic Journal*, March 1989, *99*, 116–31.

Basalla, George, *The Evolution of Technology*, Cambridge: Cambridge University Press, 1988.

David, Paul A., "Path Dependence: Putting the Past into the Future of Economics," unpublished paper, Stanford University, 1988.

Day, Richard H., "The General Theory of Disequilibrium Economics and of Economic Evolution," in D. Batten et al., eds., *Economic Evolution and Structural Adjustment*, Berlin: Springer, 1987.

Gerschenkron, Alexander, "On the Concept of Continuity in History," in *Continuity in History and Other Essays*, Cambridge: Belknap, 1968.

Goldschmidt, Richard, *The Material Basis of Evolution*, New Haven: Yale University Press, 1940.

Gould, Stephen Jay, "The Uses of Heresy," An Introduction to the Reprint of R. Goldschmidt's *The Material Basis of Evolution*, New Haven: Yale University Press, 1982.

Headrick, Daniel R, *The Invisible Weapon: Telecommunications and International Politics, 1851–1945*, New York: Oxford University Press, 1990.

Jones, Eric L., *Growth Recurring: Economic Change in World History*, Oxford: Clarendon Press, 1988.

Lewontin, Richard C., *Human Diversity*, New York: Scientific American Books, 1982.

Maynard Smith, John, *Evolution and the Theory of Games*, Cambridge: Cambridge University Press, 1982.

Mirowski, Philip, *More Heat than Light*, New York: Cambridge University Press, 1990.

Mokyr, Joel, "Evolutionary Biology, Technological Change, and Economic History," unpublished manuscript, Northwestern University, 1989.

______, (1990a) "Was There a British Industrial Evolution?," in his *The Vital One: Essays Presented to Jonathan R. T. Hughes*, Greenwich: JAI Press, 1990.

______, (1990b) *The Lever of Riches: Technological Creativity and Economic Progress*, New York: Oxford University Press, 1990.

Olson, Mancur, *The Rise and Decline of Nations*, New Haven: Yale University Press, 1982.

Sahal, Devendra, *Patterns of Technological Innovation*, Reading: Addison Wesley, 1981.

The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox

By Paul A. David*

Many observers of recent trends in the industrialized economies of the West have been perplexed by the conjecture of rapid technological innovation with disappointingly slow gains in measured productivity. A generation of economists who were brought up to identify increases in total factor productivity indexes with "technical progress" has found it quite paradoxical for the growth accountants' residual measure of "the advance of knowledge" to have vanished at the very same time that a wave of major innovations was appearing—in microelectronics, in communications technologies based on lasers and fiber optics, in composite materials, and in biotechnology. Disappointments with "the computer revolution" and the newly dawned "information age" in this regard have been keenly felt. Indeed, the notion that there is something anomalous about the prevailing state of affairs has drawn much of its appeal from the apparent failure of the wave of innovations based on the microprocessor and the memory chip to elicit a surge of growth in productivity from the sectors of the U.S. economy that recently have been investing so heavily in electronic data processing equipment (see, for example, Stephen Roach, 1987, 1988; Martin Baily and Robert Gordon, 1988). This latter aspect of the so-called "productivity paradox" attained popular currency in the succinct formulation attributed to Robert Solow: "We see the computers everywhere but in the productivity statistics."

If, however, we are prepared to approach the matter from the perspective afforded by the economic history of the large technical systems characteristic of network industries, and to keep in mind a time-scale appropriate for thinking about transitions from established technological regimes to their respective successor regimes, many features of the so-called productivity paradox will be found to be neither so unprecedented nor so puzzling as they might otherwise appear.

I

My aim here simply is to convince modern economic analysts (whether perplexed by the productivity slowdown, or not) of the immediate relevance of historical studies that trace the evolution of techno-economic regimes formed around general purpose engines.[1] The latter, typically, are key functional components embodied in hardware that can be applied as elements or modular units of the engineering designs developed for a wide variety of specific operations or processes. Accordingly, they are found ubiquitously distributed throughout such systems when the latter have attained their mature, fully elaborated state. James Watt's (separate condenser) steam engine design springs to mind readily as an example of an innovation that fulfilled this technological role in the first industrial revolution. My particular line of argument will be better served, however, by directing notice to the parallel between the modern computer and another general purpose engine, one that figured prominently in what sometimes is called the "second Industrial Revolution"—namely, the electric dynamo. (But, see also Herbert Simon, 1986.)

Although the analogy between information technology and electrical technology would have many limitations if taken very literally, it proves illuminating nonetheless. Computer and dynamo each form the nodal elements of physically distributed (transmission) networks. Both occupy key positions in a web of strongly complementary technical relationships that give rise to "network externality effects" of various kinds, and so make issues of compatibility standardization important for business strategy and public policy (see my 1987 paper and my paper with Julie Bunn, 1988). In both instances, we can recognize the emergence of an extended trajectory of incremental technical improvements, the gradual and protracted process of diffusion into widespread use, and the confluence with other streams of technological innovation, all of which are interdependent features of the dynamic process through which a general purpose engine acquires a broad domain of specific applications (see Timothy Bresnahan and Manuel Trajtenberg, 1989). Moreover, each of the principal empirical phenomena that make up modern perceptions of a productivity paradox had its striking historical precedent in the conditions that obtained a little less than a century ago in the industrialized West, including the pronounced slowdown in industrial and aggregate productivity growth experienced during the 1890-1913 era by the two leading industrial countries, Britain and the United States (see my 1989 paper, pp. 12–15, for details). In 1900, contemporary observers well might have remarked that the electric dynamos were to be seen "everywhere but in the productivity statistics!"

---

*Department of Economics, Encina Hall, Stanford University, Stanford, CA 94305. Discussions with Paul Rhode were particularly helpful early in the research. I am grateful for comments from Steve Broadberry, Jonathan Cave, Nick Crafts, among the participants in the Economic History Summer Workshop held at Warwick University, July 10–28, 1989; from Timothy Taylor; and from Shane Greenstein, Avner Greif, Edward Steinmueller, and other participants in the Technology and Productivity Workshop at Stanford, October 1989.

[1] This paper draws upon material developed in a longer work—my 1989 paper.

II

At the turn of the century, farsighted engineers already had envisaged profound transformations that electrification would bring to factories, stores, and homes. But the materialization of such visions hardly was imminent. In 1899 in the United States, electric lighting was being used in a mere 3 percent of all residences (and in only 8 percent of urban dwelling units); the horsepower capacity of all (primary and secondary) electric motors installed in manufacturing establishments in the country represented less than 5 percent of factory mechanical drive. It would take another two decades, roughly speaking, for these aggregate measures of the extent of electrification to attain the 50 percent diffusion level (see my 1989 paper, Table 3, for estimates and sources). It may be remarked that, in 1900, an observer of the progress of the "Electrical Age" stood as far distant in time from the introduction of the carbon filament incandescent lamp by Edison, and Swann (1879), and of the Edison central generating station in New York and London (1881), as today we stand from comparable "breakthrough" events in the computer revolution: the introduction of the 1043 byte memory chip (1969) and the silicon microprocessor (1970) by Intel. Although the pace of the computer's diffusion in the business and public sectors of the industrialized societies during the past two decades has been faster than that recorded for the dynamo during its comparable early phase of adoption, it has been estimated that only 10 percent of the world's 50 million business enterprises today are using computers, and only 2 percent of the world's business information has been digitized (see Peter Lewis, 1989).

The history of electrification after 1900 (see I. C. R. Byatt, 1979; Thomas Hughes, 1983; Ryoshin Minami, 1987) lends considerable plausibility to the "regime transition thesis" of Christopher Freeman and Carlotta Perez (1990). They suggest that productivity growth has been sluggish, and very well might remain so because the emergence and elaboration of a new techno-economic regime based on computer and communications innovations (supplanting the mature, ossified Fordist regime of mass production) will, more than likely, be a protracted and historically contingent affair.

Certainly, the transformation of industrial processes by the new electric power technology was a long-delayed and far from automatic business. It did not acquire real momentum in the United States until after 1914–17, when regulated regional utility rates for electricity were lowered substantially in relationship to the general price level (see my 1989 paper: Table 4, Fig. 14), and central station generating capacity came to predominate over generating capacity in *isolated* industrial plants. Furthermore, factory electrification did not reach full fruition in its technical development nor have an impact on productivity growth in manufacturing before the early 1920s. At that time only slightly more than half of factory mechanical drive capacity had been electrified. (On the significance for derived productivity growth of attaining 50 percent diffusion, see my 1989 paper, Appendix A.) This was four decades after the first central power station opened for business.

The proximate source of the delay in the exploitation of the productivity improvement potential incipient in the dynamo revolution was, in large part, the slow pace of factory electrification. The latter, in turn, was attributable to the unprofitability of replacing still serviceable manufacturing plants embodying production technologies adapted to the old regime of mechanical power derived from water and steam. Thus, it was the American industries that were enjoying the most rapid expansion in the early twentieth century (tobacco, fabricated metals, transportation equipment, and electrical machinery itself) that afforded greatest immediate scope for the construction of new, electrified plants along the lines recommended by progressive industrial engineers (see Richard DuBoff, 1979, p. 142; and Minami, pp. 138–41). More widespread opportunities to embody best-practice manufacturing applications of electric power awaited the further physical depreciation of durable factory structures, the locational obsolescence of older-vintage industrial plants sited in urban core areas, and, ultimately, the development of a general fixed capital formation boom in the expansionary macroeconomic climate of the 1920s.

The persistence of durable industrial facilities embodying older power generation and transmission equipment had further consequences that are worth noticing. During the phase of the U.S. factory electrification movement extending from the mid-1890s to the eve of the 1920s, the "group drive" system of power transmission remained in vogue (see Duboff, p. 144; Warren Devine, 1983, pp. 351, 354). With this system (in which electric motors turned separate shafting sections, so that each motor would drive related groups of machines), the retrofitting of steam- or water-powered plants typically entailed adding primary electric motors to the original stock of equipment. While factory owners rationally could ignore the sunk costs of the existing power transmission apparatus, and simply calculate whether the benefits in the form of reduced power requirements and improved machine speed control justified the marginal capital expenditures required to install the group drive system, productivity accountants would have to reckon that the original belt and shaft equipment (and the primary engines that powered them) remained in place as available capacity. The effect would be to raise the capital-output ratio in manufacturing, which militated against rapid gains in total factor productivity (TFP)—especially if the energy input savings and the quality improvements from better machine control were left out of the productivity calculation.

This sort of overlaying of one technical system upon a preexisting stratum is not unusual during historical transitions from one technological paradigm to the next. Examples can be cited from the experience of the steam revolution (G. N. von Tunzelmann, 1978, pp. 142–43, 172–73). Indeed, the same phenomenon has been remarked upon recently in the case of the computer's application in numerous data processing and recording functions, where old paper-based procedures are being retained alongside the new, microelectronic-based methods—sometimes to the detriment of each system's performance (see, for example, Baily and Gordon, pp. 401–02).

Finally, it would be a mistake to suppose that large potential gains from factory electrification were obtainable from the beginning of the century onward, just because there were farsighted electrical engineers who at the time were able to envisage many sources of cost savings that would result from exploiting the flexibility of a power transmission system based on electric wires, and the efficiency of replacing the system of shafting and belts with the so-called "unit drive" system. In the latter arrangement, individual electric motors were used to run machines of all sizes (see Devine, pp. 362ff). The advantages of the unit drive for factory design turned out to extend well beyond the savings in inputs of fuel derived from eliminating the need to keep all the line shafts turning, and the greater energy efficiency achieved by reducing friction losses in transmission. Factory structures could be radically redesigned once the need for bracing (to support the heavy shafting and belt-housings for the transmission apparatus that typically was mounted overhead) had been dispensed with. This afforded 1) savings in fixed capital through lighter factory construction, and 2) further capital savings from the shift to building single-story factories, whereas formerly the aim of reducing power losses in turning very long line shafts had dictated the erection of more costly multistory structures. Single-story, linear factory layouts, in turn, permitted 3) closer attention to optimizing materials handling, and flexible reconfiguration of machine placement and handling equipment to accommodate subsequent changes in product and process designs within the new structures. Related to this, 4) the modularity of the unit drive system and the flexibility of wiring curtailed losses of production incurred during maintenance, rearrangement of production lines, and plant retrofitting; the entire power system no longer had to be shut down in order to make changes in one department or section of the mill.

Although all this was clear enough in principle, the relevant point is that its implementation on a wide scale required working out the details in the context of many kinds of new industrial facilities, in many different locales, thereby building up a cadre of experienced factory architects and electrical engineers familiar with the new approach to manufacturing. The decentralized sort of learning process that this entailed was dependent upon the volume of demand for new industrial facilities at sites that favored reliance upon purchased electricity for power. It was, moreover, inherently uncertain and slow to gain momentum, owing in part to the structure of the industry responsible for supplying the capital that embodied the new, evolving technology. For, the business of constructing factories and shops remained extremely unconcentrated, and was characterized by a high rate of turnover of firms and skilled personnel. Difficulties in internalizing and appropriating the benefits of the technical knowledge acquired in such circumstances are likely to slow experience-based learning. A theoretical analysis of an interdependent dynamic process involving diffusion and incremental innovations based upon learning-by-doing (see my paper with Trond Olsen, 1986) demonstrates that where the capital goods embodying the new technology are competitively supplied, and there are significant knowledge spillovers among the firms in the supplying industry, the resulting pace of technology adoption will be slower than is socially optimal.

III

The preceding review of the sources of "diffusion lags" bears directly on the relationship between the timing of movements in industrial productivity, and the applications found for electric power within the industrial sector. A somewhat different class of considerations also holds part of the explanation for the sluggish growth of productivity in the United States prior to the 1920s. These have to do more with the deficiencies of the conventional productivity measures, which are especially problematic in treating the new kinds of products and process applications that tend to be found for an emergent general purpose technology during the initial phases of its development. Here, too, the story of the dynamo revolution holds noteworthy precedents for some of the problems frequently mentioned today in connection with the suspected impact of the computer (see, Baily-Gordon; and Gordon-Baily, 1989): 1) unmeasured quality changes associated with the introduction of novel commodities; and 2) the particular bias of the new technology toward expanding production of categories of goods and services that previously were not being recorded in the national income accounts.

In the case of the dynamo, initial commercial applications during the 1890–1914 era were concentrated in the fields of lighting equipment and urban transit systems. Notice that qualitative characteristics such as brightness, ease of maintenance, and fire safety were especially important attributes of incandescent lighting for stores and factories, as well as for homes—the early electric lighting systems having been designed to be closely competitive with illuminating gas on a cost basis. Likewise, the contributions to the improvement in economic welfare in the form of faster trip speeds and shorter passenger waiting times afforded by electric streetcars, and later by subways (not to mention the greater residential amenities enjoyed by urban workers who were enabled to commute to the central business district from more salubrious residential neighborhoods), all remained largely uncaptured by the conventional indexes of real product and productivity.

Measurement biases of this kind persisted in the later period of factory electrification, most notably in regard to some of the indirect benefits of implementing the "unit drive" system. One of these was the improvement in machine control achieved by eliminating the problem of belt slippage and installing variable speed d.c. motors. This yielded better quality, more standardized output without commensurately increased costs (see Devine, pp. 363ff). Factory designs adapted to the unit drive system also brought improvements in working conditions and safety. Lighter, cleaner workshops were made possible by the introduction of skylights, where formerly overhead transmission apparatus had been mounted; and also by the elimination of the myriad strands of rotating belting that previously swirled dust and grease through the factory atmosphere, and, where unenclosed within safety screening, threatened to maim or kill workers who became caught up in them.

These more qualitative indirect benefits, however, came as part of a package containing other gains that, as has been seen, took the form of more readily quantifiable resource savings. Consequently, a significantly positive cross-section association can be found between the rise in the industry's TFP growth rate (adjusted for purchased energy inputs) during the 1920s, vis-à-vis the 1910s, and the proportionate increase of its installed secondary electric motor capacity between 1919 and 1929. Making use of this cross-section relationship, approximately half of the 5 percentage point acceleration recorded in the aggregate TFP growth rate of the U.S. manufacturing sector during 1919–29 (compared with 1909–19) is accounted for statistically simply by the growth in manufacturing secondary electric motor capacity during that decade (see my 1989 paper, Table 5, and pp. 26–27).

But, even that did not exhaust the full productivity ramifications of the dynamo revolution in the industrial sector during the 1920s. An important source of measured productivity gains during this era has been found to be the capital-saving effects of the technological and organizational innovations that underlay the growth of continuous process manufacturing, and the spread of continuous shift-work, most notably in the petroleum products, paper, and chemical industries (see John Lorant, 1966, chs. 3, 4, 5). Although these developments did not involve the replacement of shafts by wires, they were bound up indirectly with the new technological regime build up around the dynamo. Advances in automatic process control engineering were dependent upon use of electrical instrumentation and electro-mechanical relays. More fundamentally, electrification was a key complementary element in the foregoing innovations because pulp- and paper-making, chemical production, and petroleum refining (like the primary metals, and the stone, clay and glass industries where there were similar movements towards electrical instrumentation for process control, and greater intensity in the utilization of fixed facilities) were the branches of manufacture that made particularly heavy use of electricity for process heat.

IV

Closer study of some economic history of technology, and familiarity with the story of the dynamo revolution in particular, should help us avoid both the pitfall of undue sanguinity and the pitfall of unrealistic impatience into which current discussions of the productivity paradox seem to plunge all too frequently. Some closing words of caution are warranted, however, to guard against the dangers of embracing the historical analogy too literally.

Computers are not dynamos. The nature of man-machine interactions and the technical problems of designing efficient interfaces for humans and computers are enormously more subtle and complex than those that arose in the implementation of electric lighting and power technology. Moreover, information as an economic commodity is not like electric current. It has special attributes (lack of superadditivity and negligible marginal costs of transfer) that make direct measurement of its production and allocation very difficult and reliance upon conventional market processes very problematic. Information is different, too, in that it can give rise to "overload," a special form of congestion effect arising from inhibitions on the exercise of the option of free disposal usually presumed to characterize standard economic commodities. Negligible costs of distribution are one cause of "overload"; information transmitters are encouraged to be indiscriminate in broadcasting their output. At the user end, free disposal may be an unjustified assumption in the economic analysis of information systems, because our cultural inheritance assigns high value to (previously scarce) information, predisposing us to try screening whatever becomes available. Yet, screening is costly; while it can contribute to a risk-averse information recipient's personal welfare, the growing duplicative allocation of human resources to coping with information overload may displace activities producing commodities that are better recorded by the national income accounts.

In defense of the historical analogy drawn here, the information structures of firms (i.e., the type of data they collect and generate, the way they distribute and process it for interpretation) may be seen as direct counterparts of the physical layouts and materials flow patterns of production and transportation systems. In one sense they are, for they constitute a form of sunk costs, and the variable cost of utilizing such a structure does not rise significantly as they age. Unlike those conventional structures and equipment stocks, however, information structures per se do not automatically undergo significant physical depreciation. Although they may become economically obsolete and be scrapped on that account, one cannot depend on the mere passage of time to create occasions to radically redesign a firm's information structures and operating modes. Consequently, there is likely to be a strong inertial component in the evolution of information-intensive production organizations.

But, even these cautionary qualifications serve only to further reinforce one of the main thrusts of the dynamo analogy. They suggest the existence of special difficulties in the commercialization of novel (information) technologies that need to be overcome before the mass of information-users can benefit in their roles as producers, and do so in ways reflected by our traditional, market-oriented indicators of productivity.

REFERENCES

Baily, Martin N. and Gordon, Robert J., "The Productivity Slowdown, Measurement Issues, and the Explosion of Computer Power," *Brookings Papers on Economic Activity*, 2:1988, 347–420.

Bresnahan, Timothy F. and Trajtenberg, Manuel, "General Purpose Technologies and Aggregate Growth," Working Paper, Department of Economics, Stanford University, January 1989.

Byatt, I. C. R., *The British Electrical Industry 1875–1914: The Economic Returns to a New Technology*, Oxford: Clarendon Press, 1979.

David, Paul A., "Computer and Dynamo: The Modern Productivity Paradox in a Not-Too-Distant Mirror," Center for Economic Policy Research, No. 172, Stanford University, July 1989.

______, "Some New Standards for the Economics of Standardization in the Information Age," in P. Dasgupta and P. L. Stoneman, eds., *Economic Policy and Technological Performance*, London: Cambridge University Press, 1987, ch. 7.

______ **and Bunn, Julie A.,** "The Economics of Gateway Technologies and the Evolution of Network Industries: Lessons from Electricity Supply History," *Information Economics and Policy*, Spring 1988, *4*, 165–202.

______ **and Olsen, Trond E.**, "Equilibrium Dynamics of Diffusion when Incremental Technological Innovations are Foreseen," *Ricerche Economiche*, October-December, 1986, *40*, 738–70.

Devine, Warren, Jr., "From Shafts to Wires: Historical Perspective on Electrification," *Journal of Economic History*, June 1983, *43*, 347–72.

DuBoff, Richard, *Electrical Power in American Manufacturing 1889–1958*, New York: Arno Press, 1979.

Freeman, Christopher, and Perez, Carlotta, "The Diffusion of Technical Innovations and Changes of Techno-economic Paradigm," in F. Arcangeli et al., eds., *The Diffusion of New Technologies*, Vol. 3: *Technology Diffusion and Economic Growth: International and National Policy Perspectives*, New York: Oxford University Press, forthcoming 1990.

Gordon, Robert J. and Baily, Martin, N., "Measurement Issues and the Productivity Slowdown in Five Major Industrial Countries," OECD, Directorate of Science, Technology and Industry, Paris, June 1989.

Hughes, Thomas P., *Networks of Power: Electrification in Western Society, 1880–1930*, Johns Hopkins University Press, 1983.

Lewis, Peter H., "The Executive Computer: Can There Be Too Much Power?," *New York Times*, December 31, 1989, p. 9.

Lorant, John H., *The Role of Capital-Improving Innovations in American Manufacturing during the 1920's*, New York: Arno Press, 1966.

Minami, Ryoshin, *Power Revolution in the Industrialization of Japan: 1885–1940*, Tokyo: Kinokuniya Co., 1987.

Roach, Stephen S., "America's Technology Dilemma: A Profile of the Information Economy," *Special Economic Study—Morgan Stanley*, New York, September 22, 1987.

______, "White Collar Productivity: A Glimmer of Hope?," *Special Economic Study—Morgan Stanley*, New York, September 16, 1988.

Simon, Herbert A., "The Steam Engine and the Computer: What Makes Technology Revolutionary?," *EDUCOM Bulletin*, Spring 1986, *22*, 2–5.

von Tunzelmann, G. N., *Steam Power and British Industrialization to 1860*, Oxford: Clarendon Press, 1978.

POVERTY AND THE UNDERCLASS†

Effects of Family and Community Background on Economic Status

By
MARY CORCORAN, ROGER GORDON, DEBORAH LAREN, AND GARY SOLON*

Policy-oriented discussions of poverty in the United States repeatedly have stressed the large influence of family and community background on income status. An early example is Michael Harrington's claim:

> ...the real explanation of why the poor are where they are is that they made the mistake of being born to the wrong parents, in the wrong section of the country, in the wrong industry, or in the wrong racial or ethnic group. Once that mistake has been made, they could have been paragons of will and morality, but most of them would never even have had a chance to get out of the other America. [1962, p. 21]

More recent writings on the "underclass" also have emphasized the importance of background effects, though they have differed in their analyses of specific causal processes. Some discussions have blamed intergenerationally transmitted poverty on lack of economic resources, discrimination, and inadequate opportunities. Others have pointed to cultural and psychological deficits caused by growing up in deprived environments. Still others have focused on the role of government programs in fostering intergenerationally transmitted patterns of illegitimacy and welfare dependence. The overall importance of family and community origins, however, has been a persistent theme.

In contrast, statistical studies often have reported evidence of only modest background influences on income status. For instance, Christopher Jencks et al. concluded: "Poverty is not primarily hereditary... . Indeed, there is nearly as much economic inequality among brothers raised in the same homes as in the general population" (1972, pp. 7–8).

Subsequent sibling studies have produced a wide range of results, but many have suggested quite moderate background effects by estimating brother correlations in various income measures of about .2 or less. Similarly, Jere Behrman and Paul Taubman (1985) estimated the father-son correlation in the natural logarithm of annual earnings to be no greater than .2. On the basis of such intergenerational results, Gary Becker declared that "low earnings are not strongly transmitted from fathers to sons..." (1988, p. 10).

Are the policy-oriented discussions that have emphasized background influences unfounded, or is there something wrong with the statistical evidence? In this paper, we use a simple statistical model to explain how problems of measurement error and unrepresentative samples have caused previous statistical studies to underestimate the magnitude of background effects. We then summarize a series of our own studies, in which we have corrected for the biases in earlier studies and estimated substantially larger background effects.

---

†*Discussants*: Ronald Mincy, Urban Institute; James Smith, Rand Corporation.

*University of Michigan, Ann Arbor, MI 48109. The research summarized in this paper has been supported by a Ford Foundation grant to Michigan's Institute of Public Policy Studies, the Small Grants Program of the Institute for Research on Poverty, and a Sloan Research Fellowship awarded to Solon.

I. A Simple Model

Let y_{ij} denote the long-run economic status (say, the permanent component of log annual earnings) of the jth son from the ith family. For discussing brother correlations, it is useful to partition y_{ij} into two orthogonal components:

$$y_{ij} = a_i + u_{ij}, \tag{1}$$

where a_i is the component shared by sons from the same family and u_{ij} is the component purely idiosyncratic to son j. The population variance in y_{ij} (σ_y^2) is the sum $\sigma_a^2 + \sigma_u^2$ of the variances of the two components, and the brother correlation in long-run status is

$$\rho = \sigma_a^2/\sigma_y^2 = \sigma_a^2/\left(\sigma_a^2 + \sigma_u^2\right). \tag{2}$$

This correlation ρ, which is the proportion of the population variance in long-run status that is common to sons from the same family, is a summary measure of the proportion that is attributable to family background. It is a broad measure in that *anything* shared by brothers (not just parental characteristics, but also community characteristics such as school quality) is included in the measured background effect. On the other hand, it is narrow in that some factors often thought of as background factors (such as genetic traits not shared by brothers) are excluded. Nevertheless, ρ has considerable appeal as a rough omnibus measure of the importance of background influences, and numerous researchers have attempted to estimate it.

For discussing intergenerational correlations, it is useful to express a_i, the family component of y_{ij}, as a regression function of x_i, the long-run economic status of the father in family i:

$$a_i = \beta x_i + z_i \tag{3}$$

where the intercept is suppressed by expressing variables as deviations from generation means, and where the error term z_i represents the combined effect of all background characteristics orthogonal to the father's economic status. Then, if $\sigma_y^2 = \sigma_x^2$ (the degree of inequality is the same in both generations), the intergenerational income correlation (the correlation between y_{ij} and x_i) is simply β. Furthermore, taking the variance of both sides of equation (3) and dividing through by $\sigma_y^2 = \sigma_x^2$ yields

$$\rho = \beta^2 + \sigma_z^2/\sigma_y^2. \tag{4}$$

This expression decomposes the brother correlation ρ into a component associated with the intergenerational correlation β and a component due to background characteristics orthogonal to father's economic status.

Let us now discuss the sources of bias in previous studies. First, limitations on data availability have forced many studies to rely on peculiarly homogeneous samples. One sibling study, for example, was based on a sample of Mormon brothers in nineteenth-century Utah, and Behrman and Taubman's intergenerational study used a fathers sample of white male twins in which both members of each pair had served in the armed forces and cooperated with a succession of surveys. Studies that draw their samples from such homogeneous sets of families tend to underestimate the true sibling or intergenerational correlations for the overall U.S. population (or for any other more heterogeneous population of interest).

In the case of siblings, if the homogeneity of the sample families causes the sample variance of the family component a_i to be less than the population variance σ_a^2, but does not somehow produce a corresponding reduction in the variance of the brother-specific component u_{ij}, equation (2) makes clear that the brother correlation in the sample will understate the population value ρ. To put it another way, in a homogeneous sample in which even unrelated individuals resemble each other, the resemblance of brothers will seem less striking. Solon (1989) demonstrates that a similar result pertains to the case of intergenerational correlations.

Second, lacking direct measures of the long-run status variables y_{ij} and x_i, previous studies have used short-run proxies, such as the log of single-year measures of earnings. This use of noisy proxies leads to a downward errors-in-variables bias in the estimation of sibling or intergenerational correlations in long-run status. In the case of siblings, suppose that economic status is measured by reported status in year t:

$$y_{ijt} = y_{ij} + v_{ijt}, \tag{5}$$

where v_{ijt}, the transitory fluctuation around long-run status, arises from both true transitory variation and random measurement error. (For simplicity, this formulation abstracts from life cycle profiles and other complications that are incorporated in Solon et al., 1989.) If the transitory component v_{ijt} is uncorrelated across brothers (which turns out empirically to be approximately true), then studies that measure the brother correlation in single-year status effectively estimate not the long-run correlation ρ, but instead

$$\sigma_a^2/(\sigma_a^2 + \sigma_u^2 + \sigma_v^2) = \phi\rho \tag{6}$$

where

$$\phi = (\sigma_a^2 + \sigma_u^2)/(\sigma_a^2 + \sigma_u^2 + \sigma_v^2). \tag{7}$$

The factor of attenuation bias, ϕ, is the proportion of the population variance in single-year status that is permanent. Results from previous longitudinal studies that have decomposed the variance of log annual earnings or wages into permanent and transitory components suggest that, even in a representative sample of the U.S. population, ϕ is less than .7. In a sample of homogeneous families, in which the "signal" is likely to be reduced without a commensurate reduction in "noise," the errors-in-variables bias is probably even more severe. Again, Solon (1989) shows that a similar analysis applies to intergenerational correlations as well.

II. Empirical Findings

To achieve more accurate estimates of the importance of background influences, we have undertaken a series of new studies based on data from the *Panel Study of Income Dynamics* (*PSID*). Because the *PSID* sample is a national probability sample, it avoids the homogeneity of samples used in previous research. Furthermore, the longitudinal nature of the data from the *PSID*'s annual reinterviews makes it possible to explore the empirical importance of using short-run vs. long-run measures of income status.

Because the *PSID* has followed children from the families originally sampled in 1968 as the children have grown into adulthood, the sample now contains multiple adult siblings from the same families. In Solon et al. (1989), we exploit the longitudinal data on these siblings to generate new estimates of sibling correlations in economic status. Our estimates of the brother correlation in the log of *single-year* earnings center around .25. This figure is slightly above the central tendency of previous estimates, which is unsurprising given the heterogeneity of our sample.

We then apply analysis-of-variance techniques to the longitudinal data to separate transitory from permanent variation and estimate ρ, the brother correlation in the *permanent* component of log earnings. Our estimate of .45 is far higher than our own single-year estimates and most others in the literature. The ratio of our single-year to permanent estimates, which estimates the attenuation bias factor ϕ, lies robustly in the range .55 to .60, and is estimated with reasonably good precision. Unfortunately, a variety of estimation problems, including the small size of our siblings sample, causes the standard errors of our estimates of the sibling correlations themselves to be quite large. We therefore are unable to reach firm conclusions about the precise magnitude of overall background effects.

Nevertheless, our estimates of ϕ, along with evidence from other longitudinal studies of earnings, strongly support the hypothesis that brother correlation estimates based on single-year data are prone to dramatic underestimation of background influences on long-run earnings. Our further analyses of brothers' hourly wages and family income and sisters' family income again estimate ϕ to be much less than 1 and yield estimates of

ρ much larger than most estimates based on single-year data.

In Solon et al. (1988), we examine the sister correlation in welfare program participation. In that paper, we observe whether each sister received public assistance each year between the time of leaving home and age 27. We estimate that a woman's probability of ever participating in welfare programs during that period is .20 if her sister has *not* participated, but is .66 if her sister *has* participated. We also consider a persistent participation measure based on whether the woman reported welfare receipt in at least three different calendar years. We estimate that a woman's probability of persistent participation is .12 if her sister is not a persistent participant, but is .56 if her sister is a persistent participant. These contrasts, estimated with good precision, imply a strikingly large background effect on welfare receipt. Additional analyses indicate that welfare program participation varies considerably with race and parental welfare use, but, even after controlling for these specific background factors, the sister resemblance in welfare receipt remains very large.

Solon (1989) uses the intergenerational span of the *PSID* to estimate β, the intergenerational correlation in long-run income status. The estimates turn out to be much more reliable than our estimates of the sibling correlation ρ, partly because the available sample of father-son pairs is larger than that of sibling pairs, and partly because the longitudinal data for separating transitory from permanent variation are more extensive for fathers than for sons. The results strongly suggest that the father-son correlation in the permanent component of log earnings is at least .4, double the .2 value previously described as an upper limit by Behrman-Taubman and by Becker and Nigel Tomes (1986). These results, along with similar findings for hourly wage rates and family income, depict a much less mobile society than was portrayed in earlier studies based on homogeneous samples and short-run status measures.

As shown in equation (4), the portion of the sibling correlation ρ associated with parental income is the *squared* intergenerational correlation. Although a β of .4 is far larger than previous estimates, its square .16 almost undoubtedly comprises a minority share of the sibling correlation in long-run status. The question immediately posed is, if parental economic status cannot account for the bulk of the sibling correlation, which other background characteristics do?

In Corcoran et al. (1989), we attempt to answer this question by estimating regression relationships between men's log earnings (as well as hourly wages, hours of work, and family income) and a wide variety of family and community background characteristics. The family background variables include race, religion, several variables describing family structure and size, parents' years of education, and a series of family employment and income variables. The community background variables include region, city size, and zip code-level statistics on median family income, male unemployment rate, percentage of families that are female headed with children, and percentage of families receiving public assistance in the men's communities of origin.

Our results continue to show strong influences of parental income, including especially large disadvantages for men raised in poverty. Even after controlling for parental income, we find substantial disadvantages for blacks and those from more welfare-dependent families or communities. The separate effect of parental education is negligible, and the zip code-level community variables aside from percentage on welfare also show little association with men's economic outcomes.

Interpretation of these findings is complicated by the presence of errors-in-variables and omitted-variables biases. For example, the small estimated community effects might be due to the crudeness of zip code-level statistics as measures of community characteristics. It is conceivable that, if community characteristics were measured in finer geographic detail, community effects might become more apparent. Another example is our finding that the presence of either family welfare receipt or high community welfare program participation is associated with negative outcomes. This negative association is consistent with Charles Murray's (1984) and Lawrence Mead's (1986) hypothesis that family or community welfare dependence sets a poor example for children's self-reliance and consequently impedes children's later achievement. On the other hand, although their hypothesis emphasizes adverse effects on work effort, we find the least evidence of a negative association for the hours-of-work outcome variable. This anomaly suggests that the observed negative associations might reflect other omitted factors. One possibility is that the community welfare variable is effectively a proxy for the "social isolation" of underclass communities emphasized by William Wilson (1987). Another possibility is that various unmeasured parental characteristics, perhaps including self-reliance, are correlated with both parental welfare use and children's outcomes. The difference between these stories and that of Murray and Mead is that the crucial background traits might be passed on to children, through nature or nurture, even if welfare programs did not exist.

III. Summary

Because of errors-in-variables problems and reliance on unrepresentatively homogeneous samples, previous statistical studies have substantially underestimated the influence of family and community background on economic status. Our own studies, designed to be less susceptible to these biases, have estimated sizable background effects on earnings, hourly wages, family income, and welfare program participation. We have found parental income (especially parental poverty), race, and parental and community welfare use to be especially strongly associated with children's economic outcomes. These findings seem broadly consistent with underclass analyses that emphasize the importance of family and community influences.

Identifying the particular causal processes underlying the estimated background effects, however, is exceedingly difficult. It is unclear to what extent sibling correlations in economic status reflect the effects of nature versus nurture and to what extent nurture effects operate through family or community influences on values or through "investments" in children's income-generating abilities. It is equally unclear to what extent these processes are shaped by discrimination and other barriers to opportunity and to what extent they can or should be modified by government policy. These questions comprise a formidable, but important agenda for continuing research.

REFERENCES

Becker, Gary S., "Family Economics and Macro Behavior," *American Economic Review*, March 1988, *78*, 1–13.

______ **and Tomes, Nigel,** "Human Capital and the Rise and Fall of Families," *Journal of Labor Economics*, July 1986, *4*, S1–39.

Behrman, Jere and Taubman, Paul, "Intergenerational Earnings Mobility in the United States," *Review of Economics and Statistics*, February 1985, *67*, 144–51.

Corcoran, Mary et al., "Effects of Family and Community Background on Men's Economic Status," NBER Working Paper No. 2896, March 1989.

Harrington, Michael, *The Other America*, Baltimore: Penguin, 1962.

Jencks, Christopher et al., *Inequality*, New York: Basic Books, 1972.

Mead, Lawrence, *Beyond Entitlement*, New York: Free Press, 1986.

Murray, Charles, *Losing Ground*, New York: Basic Books, 1984.

Solon, Gary, "Intergenerational Income Mobility in the United States," Discussion Paper No. 894-89, Institute for Research on Poverty, July 1989.

______ **et al.,** "Sibling and Intergenerational Correlations in Welfare Program Participation," *Journal of Human Resources*, Summer 1988, *23*, 388–96.

______, "A Longitudinal Analysis of Sibling Correlations in Economic Status," mimeo., University of Michigan, August 1989.

Wilson, William Julius, *The Truly Disadvantaged*, Chicago: University of Chicago Press, 1987.

AFDC Participation Across Generations

By PETER GOTTSCHALK*

William Julius Wilson's influential book *The Truly Disadvantaged* (1987) has sparked a vigorous debate about the existence and causes of an urban underclass. While there is still no clear agreement about definitions, the growing literature on the underclass focuses both on traits of the current generation and the traits of their parents. If an underclass exists, its current members are assumed to live in deteriorating inner cities, to experience long spells of joblessness or welfare dependency, and/or to engage in illegal and antisocial activities. Furthermore, these traits are assumed to be correlated across generations, leading to a cycle of deprivation.

In this paper, I focus on one such link across generations by examining the correlation in welfare participation of daughters and their parents. I first present the state of knowledge about the extent to which children of parents who received Aid to Families with Dependent Children (AFDC) are themselves more likely to participate in this welfare program. Second, I discuss the remaining gaps in our understanding and give several contrasting alternative explanations for the observed correlation, each with a different policy implication.

I. Degree of Intergenerational Correlation

Until recently, much of the debate about the intergenerational correlation of welfare participation had to take place in a factual vacuum. With the advent of longitudinal data sets sufficiently long to observe the adult experiences of children who grew up in households receiving assistance, it is now possible to measure the degree of intergenerational correlation of welfare participation with some accuracy.

Except for Sarah McLanahan (1988), the limited set of studies that do exist (John Antel, 1987; Greg Duncan et al., 1988; Lee Rainwater, 1987; and Gary Solon et al., 1987) all use the same methodology. They look at whether a mother received assistance at the beginning of the panel, and whether the daughter received welfare after she left home. All these studies find that the probability of a daughter receiving assistance is low during the relatively short period covered, but that the probability that the daughter participated is indeed higher if the mother participated in AFDC.

These studies may, however, overstate the degree of intergenerational correlation. Since all daughters are included, whether their parents were eligible for AFDC or not, some of the positive correlation found between mothers' and daughters' welfare experiences may reflect the fact that daughters of high-income parents are unlikely to have sufficiently low income to be eligible for AFDC. Put in popular terms, the policy relevant question would seem to be whether a daughter growing up in a household in which the mother is eligible for AFDC, but does not accept assistance (because the benefits are too small, the stigma costs are too high or the case worker discourages the mother from accepting assistance) is as likely to receive assistance as a daughter growing up in an eligible household that does accept aid.

In order to explore the hypothesis that participation per se matters, one can compare only daughters of parents who were eligible for AFDC. Table 1 presents data on daughters' childbearing and AFDC status for respondents cross classified by their parents' AFDC eligibility and participation sta-

---

*Professor of Economics, Boston College, Chestnut Hill, MA 02156, and Project Associate at the Institute for Research on Poverty. This project was funded under a grant from the U.S. Department of Health and Human Services/Office of the Assistant Secretary for Planning and Evaluation/Income Security Policy to the Institute for Research on Poverty, University of Wisconsin-Madison. Michael Gallaher provided exceptional research assistance. All opinions presented in this paper are my own.

TABLE 1—DISTRIBUTION OF DAUGHTERS AT THE END OF THE SURVEY PERIOD BY PARENTHOOD AND AFDC STATUS CROSS CLASSIFIED BY PARENT'S AFDC ELIGIBILITY AND PARTICIPATION

	Parent Received AFDC	Parent Did Not Receive AFDC	
		Eligible	Ineligible
	(1)	(2)	(3)
Characteristics of Parents			
Race			
Black	.416	.273	.156
White	.464	.507	.704
Hispanic	.119	.219	.139
Income			
AFDC Income[a]	$1,804	0	0
Total Income[a]	$17,161	$19,088	$34,566
No Male Present	.494	.441	.184
Mother's Education	10.1	11.2	11.5
Urban Resident[b]	.814	.707	.760
Disadvantaged Students[c]	.314	.219	.192
Distribution of Daughters by Presence of Child and AFDC Status			
Had Child	.523	.315	.276
Did Not Receive AFDC	.234	.197	.227
Received AFDC	.289	.118	.049
Childless	.477	.685	.724
Conditional Prob Receiving AFDC If Have Child	.553	.374	.176
No. of Observations	533	388	1444

[a] Average over all years.
[b] Respondent's residence in 1979.
[c] Percent in respondent's school at age 14.

tus.[1] Column 1 presents data for daughters whose parents received AFDC at some point during the survey. Column 2 is for daughters whose parents were eligible for AFDC in at least one year, but did not participate, while column 3 is for daughters with parents who were never eligible.

Table 1 shows the characteristics of the household in which the daughter grew up. The data clearly show that daughters in households that received AFDC (col. 1) had different economic and demographic characteristics than those in households that did not receive AFDC (cols. 2 and 3). Daughters growing up in households that received assistance were disproportionately nonwhite and came from more disadvantaged backgrounds, as measured either by family income, mother's education, or the proportion of disadvantaged students in the respondent's school.

Table 1 also shows daughters' childbearing and AFDC status at the end of the survey. Since columns 1 and 2 only include eligible parents, the differences in the two columns reflect factors associated with participation and not eligibility (the fact that many high-income parents were not eligible and, therefore, did not participate does not affect these two columns). Comparing columns 1 and 2 shows that, while the probability of having a child and not receiving AFDC differed only marginally between daughters whose parents participated and those who did not (.198 vs. .234), the probability of having a child and receiving assistance is substantially higher if the parent did participate in AFDC (.289 vs. .118). The result of the higher probability of having a child and receiving aid is that daughters of parents who participated not only have a higher overall probability of giving birth by the end of the survey (.523 vs. .315), but more than one-half of these young mothers also received AFDC (.533). In comparison, the conditional probability of the daughter receiving assistance, given that she had a child, is only .374 if her (eligible) parents didn't receive AFDC.

Comparing columns 2 and 3 shows that including parents ineligible for aid confounds the effects of economic circumstances (as reflected in eligibility) and the effects of recipiency per se. Both columns include only daughters of families who did not receive AFDC, so the only difference is whether the parents were eligible. While the probability of the daughter having a child and not receiving aid by the end of the survey is actually somewhat lower for daughters of parents who were eligible and didn't participate than for ineligible parents (.198 vs. .227), the probability of the daughter having a child and receiving assistance is considerably higher if her parents were eligible (.118 vs. .049). This shows that parents' economic circumstances, as measured by eligibility, are correlated with daughters' AFDC participation, even among parents who do not participate. Therefore, studies that do not distinguish between eligible and ineligible households allow parents' participation to capture the effects of having sufficiently low income to be eligible, as well as the independent impact of parents' participation.

---

[1] Data is from the *National Longitudinal Survey of Youth*. For details on definitions and methodology, see my 1989 paper.

The data in Table 1 are only suggestive, since they only control for parent's eligibility and participation. In order to measure the net effect of participation, I estimated a discrete time, competing risk, duration model that describes the duration until a daughter has her first child and her AFDC status in the year of first birth.[2] The model includes variables describing the AFDC history of the family in which the daughter grew up, as well as the family's income, the presence of a male in the household, and a set of control variables. The results show that the daughters of eligible parents who received AFDC were more likely to have an early birth and to receive AFDC than daughters of eligible parents who did not participate in AFDC. This relationship is statistically significant for both blacks and whites, but not for Hispanics. Furthermore, the impact of parent's participation is substantially larger on the odds of the daughter having a child and receiving aid, than on the odds of having a child and not receiving aid. Thus, parent's participation is significantly correlated with both early births and participation in AFDC, even after restricting the comparison to daughters who grew up in families eligible for assistance, and after controlling for family income and the presence of a male in the household.

Note that this pure participation effect holds income constant—the comparison is between the growing up in a household that received AFDC and growing up in a household receiving the same income but from different sources. However, this may not be the appropriate comparison. If the parents do not participate in AFDC, family income declines, which raises the probability of the daughter receiving assistance. Calculations using the estimated coefficients from the duration model, however, show that the effect of the parental loss of AFDC income does not fully offset the pure participation effect.

While this research suggests that the intergenerational correlation of welfare participation is more than a statistical artifact, the evidence is consistent with several competing behavioral hypothesis. These are explored in the following section.

II. Alternative Explanations for the Correlation

There is a presumption in much of the public debate over welfare policy that the observed intergenerational correlation reflects a causal link between the behaviors of parents and their children. The links are, however, seldom spelled out. Since the policy implications of the observed correlation across generations depend on the underlying cause, it is useful to sort out three conceptually different reasons why such a correlation might exist.

Some authors, such as Charles Murray (1984), have stressed the transmission of preferences as an explanation for multiple generations of welfare participants. In brief, this argument assumes that the correlation in behavior reflects correlation in tastes. If parents' participation reflects low stigma, and if these attitudes are passed on to children, then the children of welfare participants will be more likely to participate as adults. Furthermore, if participation by the child is undesirable, as is assumed in much of this literature, then the observed intergenerational correlation in participation is an indication of a social problem.

The policy implications of this explanation are, however, not clear since reducing the parent's participation will have no affect on the child's participation if the cause of the correlation is transmission of preferences —the tastes will be transmitted whether the parent participates or not. In terms familiar to economists, if the intergenerational correlation reflects heterogeneity and not state dependence, then altering the parent's state will have no effect on the child's outcome.

---

[2] See my earlier paper for details.

The second explanation for the observed intergenerational correlation in participation focus on correlation across generations in constraints, not tastes. Since AFDC can only be received if the family's income falls under the break-even income, and if its assets are sufficiently low, the observed correlation between the parent's AFDC status and the child's outcome may be nothing more than evidence of intergenerational correlation of unmeasured constraints—poor mothers are more likely to receive AFDC and their children are more likely to have sufficiently low income and assets to make them eligible for aid. The intergenerational correlation in participation is again spurious, but in this case it is constraints, not tastes, that cause the correlation.[3]

If the intergenerational correlation in participation reflects correlation in income, and if children's earnings capacity depend on their parents' access to resources, then the policy implications are very different. Breaking the cycle of dependency by making it more difficult for mothers to receive aid will not solve the child's problem. In fact, if lack of income is the cause of the intergenerational correlation in welfare participation, then higher benefits or additional programs to raise the earnings capacity of the children may lower the correlation. This is, in fact, the explicit rational for several programs that provide health and nutrition to young mothers, and for programs, such as Head Start, that compensate for the parent's disadvantage in providing human capital.

Correlated tastes and correlation incomes induce a spurious correlation between generations. A causal connection exists only if the parent's AFDC status per se alters the child's outcome. Only in this case of true state dependence would the child's outcome be affected by changing the parent's eligibility for assistance. But, why would the source of the parental income have an impact, if the two families were similar in all respects?

A simple job-matching model can give rise to such state dependence. The same mechanism that links occupations across generations may be at work in causing welfare participation to be correlated across generations. Consider young adults who must chose among jobs that have outcomes that can only be measured imperfectly until the job is experienced. Furthermore, assume that it is costly to try out a job, either because of the opportunity cost of time or the cost of having a checkered job history. The choice facing the young adult is which job to sample. The probability of accepting a job is the product of the probability of paying the price to sample that job multiplied by the conditional probability of accepting the job, given that it was sampled. Suppose that the tastes of children are independent of their parents, so the conditional probability of the child accepting the job once the job has been experienced is independent of the parent's job. However, if the price of obtaining information about the parent's job is lower than the price of obtaining information about other jobs (through supper conversations and other informal networks), then the offspring will be more likely to consider their parents' jobs, in the sense of deciding whether that job meets their minimum standard. Thus, even if the conditional probability of taking the job once it has been considered is independent of the parent's occupation, the unconditional probability is indeed higher that the offspring will have same occupation as the parent—the same proportion of those who try the job will find it acceptable, but children will be more likely to try their parent's job because the price of gaining information is lower. This will lead to intergenerational correlation in occupations.

If welfare participation is viewed as one of many options that any young adult could consider, then intergenerational correlation of welfare participation may be nothing more than a special case of intergenerational correlation of occupations—just as the children of lawyers are more likely to become lawyers, the children of welfare participants are more likely to become participants because the cost of information is lower.[4]

---

[3] Controlling for eligibility and controlling for income among the eligible population partially eliminates the effects of parents' constraints. Unobservable constraints may, however, still be correlated.

If the parent's participation matters per se, then policymakers will want to consider the intergenerational consequences of alternative policies. Raising AFDC benefits will increase the resources available for human capital development, which may lessen the need for assistance of the next generation. However, the increased welfare participation of the current generation may lead to higher participation of the following generation. In contrast, raising family income through the tax system, or other alternatives to welfare, will make more resources available to the family without inducing future participation by the child.

---

[4] This observation was first suggested by Eugene Smolensky.

REFERENCES

Antel, John, "The Inter-generational Transfer of Welfare Dependency," mimeo., University of Houston, 1987.

Duncan, Greg, Hill, Martha and Hoffman, Saul, "Welfare Dependence Within and Across Generations," *Science*, January 1988, *239*, 467–71.

Gottschalk, Peter, "The Intergenerational Transmission of Welfare Participation: Facts and Possible Causes," mimeo., Boston College, 1989.

McLanahan, Sarah, "Family Structure and Dependency: Early Transitions to Female Household Headship," *Demography*, February 1988, *25*, 1–16.

Murray, Charles, *Losing Ground*, New York: Basic Books, 1984.

Rainwater, Lee, "Class, Culture, Poverty and Welfare," Center for Human Resources, Heller Graduate School, 1987.

Solon, Gary et al., "Sibling and Intergenerational Correlations in Welfare Program Participation," mimeo., University of Michigan, 1987.

Wilson, William, J., *The Truly Disadvantaged*, Chicago: Chicago University Press, 1987.

Welfare and Work in Chicago's Inner City

By MARTA TIENDA*

Two salient themes recurring in recent writings about the urban underclass are welfare dependence and detachment from the labor market. What distinguishes the urban underclass from the poor in general is the *persistence* of these behaviors in the context of social isolation from mainstream institutions (Sara McLanahan and Irwin Garfinkel, 1989). Another theme recurring in discussions about persisting and concentrated poverty is that it is predominantly a minority phenomenon, if not one exclusive to blacks. Unfortunately, the limited availability of data to examine the prevalence of chronic poverty, economic dependence, and joblessness among nonblack minorities precludes verification of claims that Latinos, notably those of Puerto Rican origin, may be joining the urban underclass.

I address these issues by examining how employment facilitates exits from AFDC spells for a unique sample of mothers who resided in poor ethnic neighborhoods. Specific questions addressed are: 1) What is the prevalence of short vs. long spells of AFDC recipiency among inner city residents, and how does this distribution differ among black, Latino, and white mothers? 2) How does employment influence the probability of exiting AFDC, and is the influence of work on departures from AFDC uniform for black, Latino, and white mothers?

I. Analytical Framework

Although currently there is no formal "theory" of welfare behavior (Robert Plotnick, 1983), a large body of empirical research has identified a set of factors that expose individuals to the risk of poverty, and other risks associated with the duration of welfare spells. Prior studies show that the longest spells correspond to women who are not high school graduates, who bear children out of wedlock, who are black, and who have no labor market experience (June O'Neill et al., 1987). Women living in poor urban neighborhoods also experience high risks of receiving welfare (Kathleen Harris, 1989).

Most studies have identified race as a correlate of welfare receipt and duration (but see Plotnick), yet race effects have not been *explained*. Two possibilities exist. One (essentially an argument about composition) is that racial groups differ in their access to resources conducive to economic self-sufficiency. This explanation is not compelling because most empirical analyses account for resource stocks that determine the risk of welfare receipt. An alternative explanation is one of heterogeneity, namely that racial groups differ in their disposition toward welfare participation in ways not observed by researchers. These conceptions of race effects have implications for the modeling strategy adopted.

A methodological issue raised by prior research concerns the time metric used to measure duration. Plotnick identified numerous problems associated with measuring welfare duration on an annualized basis. Not only are there reasons for expecting welfare transitions to occur over periods shorter than years, but the measurement of duration in year intervals may distort empirical results. Finally, an examination of how marriage, work, or other changes in states influence welfare exit probabilities inevitably raises questions of endogeneity. To avoid endogeneity biases in estimating the effect of work on the duration of welfare spells, I distinguish between work experience acquired *prior* to the onset of a spell and work performed *during* a spell of AFDC. The latter permits changes in employment status to occur each year of a given spell.

---

*Professor of Sociology, University of Chicago, Chicago, IL., 60637. This research was supported with funds provided by grants from the Rockefeller Foundation and the Institute for Research on Poverty to the University of Chicago. I gratefully acknowledge programming and research assistance from Haya Stier and helpful advice from Seth Sanders.

Model Specification. Empirical estimates of the duration of welfare spells included fixed and time-varying covariates denoting respondent's family and individual background, attributes at the beginning of the spell, and the social environment in which welfare behavior occurs. Spells of AFDC, defined as any series of consecutive months in which respondents reported being on a grant, were the unit of observation. My hypothesis of race and ethnic differences in the process of exiting welfare required allowance for nonproportional effects of explanatory factors on the hazard of exiting AFDC. This was achieved by stratifying according to race, the variable assumed to have nonproportional effects (Paul Allison, 1984).

Three variables in the model portray the socioeconomic circumstances of respondents' family of orientation: experience in broken families; receipt of AFDC during childhood; and mother's education. Respondent's attributes that influence duration include age at first birth and nativity. Other fixed covariates were measured at the time welfare spells began. These include respondent's age, the presence of children, whether recipients had graduated from high school, their marital status, and the cumulative years of work experience. A variable denoting spell number differentiates initial from return spells. Finally, the model included four covariates whose value changed over time. The primary variable of interest was whether respondents worked each year during the spell. Time-varying covariates denoting whether respondents became divorced during the spell, whether they graduated from high school during the spell, or whether they gave birth each year during the spell were introduced as controls.

Inclusion of a covariate measuring the poverty status of the neighborhood at the time of the survey, while potentially problematic, indicates whether the length of welfare spells covaries with the relative concentration of poverty. Current arguments about the pervasiveness of chronic dependence in ghetto poverty areas predict an inverse association between the neighborhood poverty rate and the probability of exiting AFDC (William Wilson, 1987).

Estimation Procedures. Exit probabilities were derived from survival analyses computed separately for black, white, Mexican, and Puerto Rican mothers who ever experienced a spell of AFDC. To focus on the importance of work as an exit path, I recomputed the survival analyses according to exit type, that is, via a job or other (primarily marriage) reasons. Proportional hazards models with time-dependent covariates were estimated to analyze what factors determine the length of welfare spells. For these analyses spell duration, measured in century months, was the dependent variable. Extensive exploratory analyses revealed that the survivor functions assumed quite different shapes for black, white, Mexican, and Puerto Rican women. Stratification by race and ethnic group captures differences in the arbitrary time function underlying the length of welfare spells for my sample.

II. Data, Sample Profile, and Estimation Procedures

The data for the empirical analysis is the *Urban Family Life Survey* (*UFLS*) conducted by NORC (William J. Wilson, Principal Investigator) during 1986 and 1987 in the City of Chicago. This survey was based on a multistage, stratified probability sample of parents aged 18–44 who resided in census tracts where 20 percent or more of families had incomes below the 1980 federal poverty line. Completed interviews were obtained from 2,490 respondents, representing a 79 percent response rate. The final sample includes 1,184 black, 365 white, 488 Mexican origin, and 453 Puerto Rican respondents.

The *UFLS* is comprehensive in its coverage of respondents' social and family background, and it includes extensive life history calendars recording birth, schooling, work, welfare, and relationship histories retrospectively. This strategy minimizes problems associated with left censoring, but increases reporting error through recall. A comparison of the demographic and educational characteristics of the Chicago sample with a recent national survey revealed no significant differences by race and ethnicity, but formidable differences in work and welfare behavior were observed.

The work and welfare event histories provide the basis for the statistical analyses, but my empirical model also includes information from the birth, marriage, and school histories. Welfare histories contain information for up to four spells of public assistance for each respondent. The work histories used a definition of steady work, defined as jobs that lasted at least 6 months. Work conducted at home, odd jobs, or part-time work were included in the work history if the activity lasted 6 months or more.

Sample Profile. A descriptive profile revealed appreciable ethnic differences in welfare utilization. Over 80 percent of black mothers and three-fourths of Puerto Rican mothers reported at least one welfare spell, compared to less than 30 percent of Mexican mothers. White mothers were intermediate between these extremes, with just under half reporting at least one bout of dependence. The low rates of welfare use among Mexican mothers reflects the high immigrant composition of this group. Foreign birth reduces welfare participation for two reasons. First, undocumented women are ineligible, and disinclined to apply for assistance to avoid detection. Second, the absence of public assistance programs in Mexico may dampen their demand for assistance. By contrast, dependence is widespread in Puerto Rico, hence island birth insulates Puerto Rican mothers neither from the idea of welfare, nor the prospect of its sustained use.

Welfare recipients were considerably more likely to report never having had a steady job. Mexican mothers were the exception in that low rates of labor force participation characterize welfare participants and nonparticipants. Mother-only families were more pervasive among those who ever received public assistance, even though the incidence of single headship was greatest among blacks and Puerto Ricans. Similarly, while larger shares of mothers who received public aid as adults also reported having received assistance as children, exposure to welfare during childhood was more pervasive among blacks and Puerto Ricans compared to whites and Mexicans.

III. Empirical Results

My statistical analysis is restricted to mothers, and for the duration analyses, to those who ever received AFDC. The conditional AFDC exit probabilities for mothers who resided in poor Chicago neighborhoods revealed sharp race and ethnic differences. Mexican AFDC mothers stood at one extreme, with an exit probability of just under 25 percent the first year on a grant. Puerto Rican mothers marked the opposite extreme with less than a 5 percent probability of exiting AFDC during the first year. Although exit rates decline with duration, the drop is not monotonic. Black exit rates exhibit the greatest monotonicity by duration, but only through the first 7 years of receipt. Thereafter, exit probabilities, while generally low, fluctuated in both directions, a pattern repeated among Puerto Rican and white mothers.

A comparison of exits by type revealed substantially higher exit probabilities through work vs. other (primarily marriages) means for both Puerto Rican and black mothers. Despite the higher probability of exiting through work at any duration, for neither blacks nor Puerto Ricans did employment constitute the modal route out of dependency. Approximately 10 percent of exits from AFDC occurred through work, while the remainder resulted through other means. Unfortunately, owing to small sample sizes, it was not possible to disaggregate exits by type for Mexican and white mothers.

Another striking result concerns the implications about prolonged durations on welfare. Contrary to findings based on national samples, mothers residing in Chicago's inner city exhibit chronic dependence. The hazard rates for black mothers imply a 70 percent probability that a spell of AFDC would last 5 years, and a 60 percent probability of a 10-year duration. The implied spell durations for Puerto Ricans were similar, but longer still: for them, there was a 77 percent probability that a spell of AFDC would last 5 years, and a 63 percent probability that it exceeded 10 years. Exit rates for Mexican mothers, on the other hand, were quite similar to those generated from national samples. Half of Mexican mothers exited AFDC by the third year of a spell; however, for those remaining more than 3 years, dependence is very difficult to break. Findings for Mexicans must be interpreted with some caution because of the unrepresentative immigrant composition of this group, and because of the small number of spells used in estimation.

The view of prolonged dependence found additional support from descriptive statistics. For the sample as a whole, completed AFDC spells ranged from a few months to 23 years, and incomplete spells lasted up to 26 years! The mean duration for all spells and completed spells, respectively, was 5 and 3 years. A disaggregation of spells according to race and ethnicity showed the average duration of completed spells for Mexican mothers to be less than a year vs. 2.2 years for all spells. For them, only 14 percent of all spells in progress exceeded 4 years duration, and only 4 percent of completed spells lasted more than 2 years. A very different profile emerged for black and Puerto Rican mothers: completed spells averaged 3 and 3.4 years, respectively, while those in progress averaged 5.2 and 5.4 years, respectively. Moreover the distribution of completed spells extended to 16 years for Puerto Ricans and 23 years for blacks. Of spells in progress, the highest duration (26 years) corresponded to Puerto Ricans, not blacks!

On balance, these data provide strong evidence of chronic dependence in Chicago's inner city. Although 40 percent of all welfare spells (including those in progress) were 2 years or less, roughly 18 percent of all spells lasted 10 or more years. Long durations were particularly common among black and Puerto Rican mothers, but not Mexican and white mothers. This information, while helpful for identifying groups at risk of chronic dependence, does not address the circumstances that determine the duration of spells. For this, the results from the hazards regressions are instructive.

Two hazard regressions were estimated: Model 1 included covariates fixed in time, while model 2 added covariates whose values changed during the spell. Most variables behaved as hypothesized, except that their relative importance differed by race and ethnicity. This is a noteworthy finding because many prior studies have postulated race effects on welfare duration to be additive, a formulation challenged by the Chicago inner city sample. Family background effects on welfare behavior emerged only for blacks and Puerto Ricans, but the transmission mechanisms apparently differ. For blacks, having a mother with a high school education increased the conditional probability of exiting welfare by roughly 50 percent relative to those with less-educated mothers. A strong intergenerational effect of headship surfaced for Puerto Rican mothers, but it was eliminated by the introduction of time-varying covariates.

Native-born Mexican mothers experienced longer welfare spells than those born abroad, but island and mainland birth did not affect the spell duration for Puerto Ricans. The presence of children influenced welfare exits only for black mothers. Residence in poor neighborhoods (a proxy for social environment and opportunities) was associated with lower AFDC exit probabilities for white and black mothers, but not Latinas. However, for black women this effect emerged only for areas with 40 percent poverty rates, while for whites, a 30 percent neighborhood poverty rate was sufficient to prolong welfare spells. Among the resources women can draw on to acquire self-sufficiency, high school credentials increased exit probabilities only for black and Mexican women, while prior work experience shortened welfare spells for white and Puerto Rican mothers. These results tentatively suggest that routes to economic independence may be enhanced through different emphases for these social groups.

Several interesting results emerged from the models with time-varying covariates. First, and consistent with prior research (for example, Harris), high school graduation increased the probability of exiting dependency, but only for black mothers. Over 10 percent of black mothers graduated from high school during an AFDC spell, while less than 5 percent of similarly situated Puerto Rican and white mothers did. Dependent Mexican mothers tend not to invest in further schooling. Second, the event of divorce or separation significantly reduced the probability of exiting welfare for all groups, while being married at the start of a spell increased the likelihood of exiting in a given year. Surprisingly, additional births during the spell lowered exit probabilities only for white women, but this effect was mediated through recidivism. That is, the exit probabilities of white recipients in first spells were greater than those of their (statistical) counterparts in second spells once changes in fertility were modeled. Finally, changes in work status produced the expected positive effect on welfare exits, but these attained statistical significance only for black and Puerto Rican women.

V. Conclusions

The empirical results reveal different exit paths from dependency among black, white, and Latina mothers. Not only do the incidence of welfare receipt and welfare duration distributions vary by race and ethnicity, but the constellation of personal and social circumstances that eventuate in chronic dependence also differ. Yet to be explained by future researchers are the nonproportional effects of key explanatory variables on the duration of welfare spells. Nevertheless, based on the duration of welfare spells, I conclude that Puerto Ricans are part of the urban underclass.

Results showing that residence in concentrated poverty neighborhoods is associated with prolonged dependence suggests that policy initiatives designed to reduce the length of welfare spells should focus both on individuals and the environments that sustain dependent behavior. Just how policy should be modified to take into account the influence of neighborhood effects is a more difficult question, but its answer depends on how neighborhood effects are believed to operate. This question warrants further research.

I must emphasize the tentative character of these results because I have not made any attempt to adjust for unobserved heterogeneity. However, I stand behind my conclusion that we can learn more about the processes shaping the length of welfare spells by acknowledging the importance of race and ethnicity in shaping different trajectories of dependence, and suggesting that public policy should similarly reflect these differences. I do not recommend race-specific policies, but suggest that race-specific emphases will enhance policy outcomes.

REFERENCES

Allison, Paul D., *Event History Analysis*, Quantitative Applications in the Social Sciences, No. 46, Beverly Hills: Sage. 1984.

Harris, Kathleen Mullan, "Teenage Mothers and Welfare Dependency: Combining Work and Welfare," paper presented at the annual meetings of the American Sociological Association, San Francisco, 1989.

McLanahan, Sara and Garfinkel, Irwin, "Single Mothers, the Underclass, and Social Policy," *Annals of the American Academy of Political and Social Science*, January 1989, *501*, 92–104.

O'Neill, June A., Bassi, Laurie J. and Wolf, Douglas A., "The Duration of Welfare Spells," *Review of Economics and Statistics*, May 1987, *69*, 241–49.

Plotnick, Robert, "Turnover in the AFDC Population: An Event History Analysis," *Journal of Human Resources*, Winter 1983, *18*, 65–81.

Wilson, William J., *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*, Chicago: University of Chicago Press, 1987.

ARE WE SAVING TOO LITTLE?[†]

International Differences in Saving

By BARRY P. BOSWORTH*

The adequacy of national saving has taken on increasing importance in an era of large external imbalances among the major industrial countries. At least among economists, current account imbalances are seen as largely the outcome of domestic patterns of saving and investment rather than trade practices. Thus, the low level of saving in the United States, rather than a failure of Japan to open its economy, is the primary cause of the U.S. current account deficit. Most of the discussion of the decline of U.S. saving has focused upon the emergence of a large government budget deficit in the 1980s, and has largely ignored an equally large decline in an already low level of private saving. At the same time, much of what we think we know about private saving behavior and the potential effectiveness of various policies to influence it are dominated by experience only within our own country. But, many of the factors in which we are most interested have exhibited very limited historical variation within the United States, and of necessity the conclusions are ambiguous.

The purpose of this note is to make use of an underutilized data source, that provides information on the saving and investment of the industrial countries within a common accounting framework, to examine the pattern of change over time, and the similarity of response to various economic disturbances in the 1970s and 1980s. The basic annual data are drawn from the System of National Accounts (SNA) submitted by the individual member countries to the OECD for the period, 1960–87. While the original analysis was undertaken for the seven largest nations plus a representative set of six smaller countries, but the summary reported here is limited to the United States, Japan, and an aggregate of nine European countries.

The data are presented within a simple accounting framework that relates the current account surplus (CA) to the difference between domestic saving (S) and domestic investment (I). Domestic saving and investment can be further subdivided into their public and private components,

$$(1)\qquad CA = Sp + Sg - Ip - Ig.$$

Because of the present focus on private saving, government saving, and investment are collapsed into a single measure of the budget deficit, net lending:

$$(2)\qquad CA = Sp - Ip + NLg.$$

The trends in the basic data are summarized in Table 1 for the United States, Japan, and the aggregate of European countries.

The basic point I would like to emphasize is that the decline in the private saving rate is not unique to the United States: private rates of saving within nearly all of the industrial countries have been drifting down for nearly two decades. In fact, the decline within the United States has generally been less than that of other countries. Instead, the contrast is that, outside the United States, the fall in rates of private saving has been exceeded by the fall in domestic rates of investment. Thus, while the current account deficits of the United States are often traced to the deficiency of national saving, the other major industrial countries are, in general,

---

[†] *Discussants*: Robert Barrow, Harvard University; Paul Romer, University of Chicago; N. Gregory Mankiw, Harvard University.

*The Brookings Institution, 1775 Massachusetts Ave., NW, Washington, D.C. 20036. A longer version of this paper containing a more detailed discussion of the data and statistical results is available upon request from the author.

TABLE 1—PRIVATE SAVING (S) AND INVESTMENT (I) AS SHARES OF NET DOMESTIC PRODUCT

	United States	Japan	Europe
1965–72			
Sp	10.5	20.1	13.5
Ip	8.7	20.7	13.3
1984–87			
Sp	8.1	12.1	11.4
Ip	6.6	8.3	6.7
Change in:			
Sp	−2.4	−8.0	−3.1
Ip	−2.1	−12.4	−6.6
Output Growth			
1960–73	4.0	9.6	4.5
1973–87	2.5	3.7	2.0
Wealth-Income Ratio			
1965–72	2.6	2.1	3.0
1984–87	3.2	3.3	6.6
Capital-Output Ratio			
1965–72	2.2	2.2	3.0
1984–87	2.6	2.2	3.4

plagued with the opposite problem: a surplus of private saving over private investment.

This tendency for private saving to exceed investment first emerged after the first oil shock in 1973 and has steadily grown in magnitude. The initial response in the 1970s was to absorb the surplus through larger government budget deficits. Efforts to promote an expansion of investment through low, and even negative, real interest rates were unsuccessful. The alternative of shipping the excess saving abroad in the form of a current account surplus was severely constrained by the need to find other countries willing to tolerate a current account deficit. At the time, the world economy was already under pressure to recycle the surpluses of the oil-producing countries. The common pattern of large budget deficits represented a sharp change from the experience of the 1960s when Japan and the European countries relied on budget surpluses to augment private saving.

In the 1980s, there has been a dramatic shift in the United States toward increased consumption, both public and private, and the emergence of a saving shortfall and a large trade deficit. Thus, other countries (particularly, Japan and Germany) could respond with a shift of their own policies, substituting current account surpluses for budget deficits. It is in this sense that today's imbalances of current accounts are a reflection of underlying changes in domestic saving and investment.

I. A Simple Perspective on Saving and Investment

In attempting to account for the shift of the saving-investment balance of Japan and Europe, people have focused attention on the evident decline in rates of investment. Yet, as I will illustrate, the surprise, from the perspective of standard economic theory, is that private rates of saving remain so high, not that investment rates are low.

Standard neoclassical growth theory emphasizes the constancy, in the absence of changes in relative prices, of the capital-output ratio. And, if investment is interpreted as an adjustment to a desired stock, the share of net investment in output should be proportionate to the rate of output growth. Similarly, most theories of consumption emphasize a target wealth-income ratio that depends on demographic and other factors, but is supposed to be invariant to the level of income. Thus, the saving rate is also expected to be proportionate to the rate of income growth.

The behavior of private saving and investment rates, therefore, should be closely linked, and changes in both should be proportionate to changes in the underlying rate of economic growth. Furthermore, it is possible to obtain crude estimates of the desired capital-output and wealth-income ratios by dividing the average saving and investment rates by the rate of income growth.

Those relationships are shown in the lower part of Table 1. The time span is divided into the period of high growth (pre-1973) and low growth (post-1973). The saving and investment rates are averages of years that exclude recessions in order to minimize the influence of short-run factors. The extent of the decline in the private saving and investment rates are shown in the fifth and sixth rows, and the shift in the balance is shown in the seventh row. In the more detailed analysis, the magnitude of decline in the investment was greater than that of saving in every country except the United States and Australia, and generally by a substantial amount.

Second, the major potential explanation of the decline, the slowing of economic growth, is measured in the eighth through tenth rows that show the rates of real GDP growth in the pre- and post-1973 periods. The slowdown is largest in percentage terms for Japan and Europe, frequently exceeding 50 percent of the pre-1973 rate.

Implied target wealth-income and capital-output ratios are calculated for each of the two subperiods by dividing the growth rate into the saving and investment rates, respectively. One interesting result is that the adjustment for differences in income growth substantially changes the relative rankings of the countries: the United States is no longer particularly low in terms of its rate of wealth accumulation relative to income, and Japan is not high.

In addition, in comparing the pre- and post-1973 period, the decline in the rates of both saving and investment is substantially less than proportionate to the slowing of output growth. Thus, the target wealth-income and capital-output ratios appear to have increased after 1973. The changes in the estimated capital-output ratios are far smaller than those for the wealth-output ratio, however; and in many cases the estimates come close to the expectation of a constant capital-output ratio. The wealth-income ratios, on the other hand, are extremely variable and imply very low marginal responses to the slowing of income.

This latter point can be made more explicit by regressing the flow rates on a 6-year moving average of real income changes for each of the original set of thirteen countries. For investment, the marginal capital-output ratios were clustered around 2.0 and were of high statistical significance. On the other hand, the marginal wealth-income ratios were positive and significant for only five of the thirteen countries, and, in every case, they were much less than the average. Output growth can explain 60–95 percent of the annual variation in investment rates, but less than 40 percent of the variation in saving rates.

The notion that investment is *not* abnormally low during the post-1973 period is given additional support by data from the OECD that shows a common pattern of rising capital-output ratios and falling rates of return on physical capital over the period as a whole. Instead, the implication is that an analytical framework that incorporates the notion of a target capital-output ratio is useful for interpreting investment behavior; but aggregate saving trends do not conform well with any theory of saving that implies a stable long-run wealth-income target.

II. An Expanded Model

The above framework may be too simple for interpreting saving behavior because it ignores a large number of factors that are stressed in the literature on saving behavior. It is possible that the depressive effects of lower income growth have been offset by changes in other factors. Thus, I extended the statistical analysis to an expanded empirical model that incorporated measures of demographics, interest rates and inflation, and Ricardo-Barro equivalence between the public and private sectors.

The measure of demographic change was constructed by reestimating a cross-national relationship that related international differences in saving to variations in income growth, the proportion of the population that is young or retired, and the proportion of males over 65 as a measure of changes in the length of the retirement period (see Martin Feldstein, 1980). The resulting coefficients on the three demographic variables were used to construct an index for each country in the sample extending over the period, 1965–87.

According to the indexes, the effect of demographics has been to increase private saving rates over the last two decades by substantial amounts, and the magnitude of change differs substantially among the countries in the sample. In general, a large decline in the number of young people has offset the rise in the proportion of the population that is retired. At the same time, an increased propensity to retire has exerted a substantial positive effect on the private saving rate of Japan and Europe. In fact, it is the change in the retirement rate, not the age distribution, that dominates the change in the indexes. Only in Sweden is the historical effect on saving negative.

The long-term bond rate (available from the OECD) was used as a measure of the nominal return on savings and inflation was measured with a 3-year weighted average change in the consumer price deflator. No adjustment is made for variations in tax rates because of the lack of available data, but I would expect that to be a relatively minor marginal effect in view of the wide variation in market interest rates.

The rate of price inflation was included as a separate variable to allow for several different potential influences on saving. First, it is an element of the real rate of interest and may provide an incentive for individuals to lower their rate of saving. The two components of the real rate may not exert equal offsetting effects on saving, however. This would be the case if individuals were subject to liquidity constraints, or limited in their ability to hedge their wealth portfolio against inflation risks. Alternatively, it has been suggested that increases in inflation may translate into increased uncertainty about the future and thus increase precautionary saving.

Third, in reporting interest income, the SNA does not adjust for the inflation component of interest income that simply compensates bondholders for the decline in the real value of their assets. Thus, inflation leads to a misstatement of private sector income that depends upon the extent to which it has a net creditor or debtor position. If we ignore changes in net foreign balances, we can adjust for this influence by multiplying the rate of inflation by the outstanding net financial liabilities of government.

Finally, Ricardo-Barro equivalence suggests that individuals undertake actions to neutralize the effects of variations in public saving in their own intertemporal spending plans. I tested for its potential relevance by including the public-sector saving rate as a determinant of the private saving rate.

While the results can only be summarized in this note, they were striking primarily for their lack of uniformity. The inclusion of factors other than income change greatly improved the overall statistical performance, but the role of individual variables differed substantially from country to country. There was some support for a demographic effect in five of the eleven countries for which data were available, but it had the wrong sign in two countries and often diverged significantly from the expected value of unity.

Interest rates were significant for a few countries, but showed a large correlation with saving rates only for Great Britain. Inflation generally had a positive net effect on saving, rather than the negative sign that would be expected if its dominant effect was operating through a real interest rate mechanism. Adjusting for the effect of inflation in measuring the income on net financial assets resulted in a significant improvement for seven of the countries, but it had a perverse effect for six countries.

There was often evidence of negative correlation between variations in government and private saving rates, and it was statistically significant in seven countries. However, there is some concern that the correlation reflects common cyclical influences on both variables because the government saving rate never approached significance if it was lagged by one period. In fact, all of the coefficient estimates were quite sensitive to minor variations in specification.

The estimation of the saving relationship within a broader model does result in a substantial increase in the statistical importance of income change in every country. However, the absolute magnitude of the marginal effect of income growth (ranging from 0.5 to 1.5) remained far below the average wealth-income ratios that are clustered about 3.0.

III. Conclusion

The United States does not stand alone in having experienced a decline in its private rates of saving. In fact, the decline has been even larger in other industrial countries, albeit they generally started from higher rates. It appears to be a general phenomenon that can be traced to the lower rates of income growth that have prevailed since the first oil shock of 1973.

Instead, the United States differs from other countries in the extent of the shortfall of private saving relative to private investment—other countries have, in general, switched to a surplus of private saving over investment. In that sense, the continuation of a large current account deficit by the United States has been far more sustainable from a global perspective than was anticipated initially. The consumption boom in the United States and the resultant trade deficit has relieved some of the problems of deficient domestic demand that plagued other industrial countries in the late 1970s and early 1980s. In effect, Americans like to consume and the Japanese like to save, and the two have struck a bargain.

In addition, the behavior of saving rates at the aggregate level does not seem consistent with many of the standard models that imply a stable long-run relationship between wealth and income. The decline in saving rates has been less than expected from models that assumed that the rate of net saving is proportionate to income growth.

The excess of domestic saving over investment in these countries might reemerge in the unlikely event that the United States should act to reduce its borrowing in world markets. It is sometimes argued that any national saving surpluses could be easily absorbed in the international capital market; but, the current market is notable for the extent to which it is a market between industrial countries only—less capital flows to the developing nations than in the 1970s.

Finally, there is some suggestion in the previous statistical results that demographic effects of aging populations (reducing private saving) may provide an automatic adjustment for these countries in the future. On the other hand, the aging of their populations will also imply a slower growth of the labor force, and thus a lower rate of investment. The net effects of demographic change on the saving-investment balance remain very uncertain. It is, however, an issue of substantial significance because of the major demographic changes that will occur in the industrial countries over the next several decades.

REFERENCES

Barro, Robert J., "Are Government Bonds New Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

Feldstein, Martin S., "International Differences in Social Security and Saving," *Journal of Public Economy*, October 1980, *14*, 225–44.

OECD, Department of Economics and Statistics, *National Accounts*, Vol. 2, data tape, 1989.

Public Debt as Private Liquidity

By MICHAEL WOODFORD*

There has been a great deal of public concern in the United States about the rapid growth of the public debt during the 1980s. Among economists, the grounds for concern most often stressed are that a higher public debt should reduce national savings (and, as a result, capital accumulation), with consequences for future income that are expected to lower welfare, at least in the long run. These consequences are predicted by what is often referred to (B. Douglas Bernheim, 1989) as the "neoclassical" model, a general equilibrium version of the "life cycle" model of consumption and savings developed by Modigliani and associates. Of course, according to the "Ricardian" view (Robert Barro, 1989), changes in the level of public debt should have no effect at all (except insofar as the higher future taxes implied by a higher public debt distort incentives), so that the concern is largely misplaced. But many economists are skeptical about the practical relevance of the Ricardian view, because of the apparent connection between government deficits and a variety of aspects of economic activity and financial market conditions (see, for example, Bernheim, and Robert Eisner, 1989). For example, the high real interest rates of the 1980s are often attributed to the rapid growth of the U.S. public debt in this period. As a result, the other predictions of the neoclassical model are widely accepted as well.

I wish to argue that the analysis provided by the neoclassical model may not be an adequate guide to policy, even if certain of its predictions are correct. Instead, I direct attention to an alternative explanation of the effects of changes in the level of public debt, which leads to very different conclusions about the welfare consequences of such policies. According to this view, "Ricardian equivalence" fails because of imperfect financial intermediation. Some economic units are liquidity constrained, which is to say that they are unable to borrow against their future income at a rate of interest as low as that at which the government borrows. Increased government borrowing can benefit such parties, insofar as they effectively receive a highly liquid asset, government debt, in exchange for giving the government an increased claim on their future income, their own claim to which represented a highly illiquid asset. A higher public debt, insofar as it implies a higher proportion of liquid assets in private sector wealth, increases the flexibility of the private sector in responding to variations in both income and spending opportunities, and so can increase economic efficiency.

There are several reasons to prefer the liquidity-constrained model to the neoclassical model as an explanation of the nonneutrality of government deficits. One is that the liquidity constraint hypothesis can explain the persistently low real returns on U.S. public debt relative to other kinds of assets.[1] Another is that the liquidity-constrained model is not vulnerable to the Barro critique, according to which altruistic bequests between generations should completely eliminate the nonneutralities associated with the neoclassical model. A third is the theoretical parsimony of an explanation of the long- and short-run effects of government deficits along essentially the same lines. In the pure neoclassical model, with efficient financial intermediation, government deficits, even

---

*University of Chicago, Chicago, IL 60637. I thank Julio Rotemberg and Larry Summers for helpful discussions, and the NSF for research support.

[1] Models with a similar structure to the one described here are used to explain both the low real return on Treasury bills, and the occurrence of long-lasting shifts in that return as a result of shifts in monetary and fiscal policies, in Javier Diaz-Gimenes and Edward Prescott (1989), and Mark Huggett (1989).

when expected to result in a permanently higher level of government debt, should have negligible effects upon spending decisions in the short run (James Poterba and Lawrence Summers, 1987), even though the long-run effects may be significant due to the eventual effect on the capital stock. As a result, liquidity constraints are often invoked to explain apparent short-run nonneutralities. It is not clear, then, why one should not also emphasize the consequences of liquidity constraints for the long-run effects (that are the main focus of current policy discussions) as well.

Finally, the liquidity-constrained model predicts that variations in the public debt should be important as such, both in the short and in the long run. As Laurence Kotlikoff (1988) has observed, in the neoclassical model, government deficits are nonneutral only insofar as they are equivalent to a transfer of wealth between generations. Accordingly, it is not the public debt as such that matters in that model, but rather the degree to which such intergenerational redistribution occurs—an appropriate measure of which would have to take into account other aspects of fiscal policy as well. Indeed, Kotlikoff argues that once these other aspects of fiscal policy, such as changes in Social Security provisions, are taken into account, the 1970s would have to be judged a period of relatively "loose" fiscal policy, and the 1980s a period of relatively "tight" fiscal policy. This means that the neoclassical model cannot be used to explain *any* of the supposed effects of high deficits in the 1980s, while the liquidity-constrained model presented here can.

I. Public Debt in a Liquidity-Constrained Economy

Consider an economy made up of two types of infinite lived households, with the number of each normalized as one. Type A households have endowment $e_1(1+g)^t$ in all even periods and $e_2(1+g)^t$ in all odd periods, while type B households have endowment $e_2(1+g)^t$ in even periods and $e_1(1+g)^t$ in odd periods, where $e_1 > e_2 \geq 0$. Both types ($i = A, B$) seek to maximize an infinite horizon objective function

$$\sum_{t=0}^{\infty} \beta^t (1+g)^t v\left(C_t^i/(1+g)^t\right),$$

where v is an increasing, strictly concave function, and where C_t^i denotes consumption in period t by each household of type i. (We may suppose that each household is an infinite lived family whose members increase at the rate g, with per capita endowment remaining constant; the family pursues a joint consumption and savings program to maximize a discounted sum of individual family members' utilities.) Total lump sum tax collections in period t will be assumed to be equally divided across the two types, in the amount $T_t/2$ per household. For simplicity, there is no government consumption. Finally, each consumer has the opportunity each period to save by holding government bonds earning a real rate of interest r, or by accumulating capital, but is unable to borrow against his future endowments. The real value of the outstanding government debt at the end of period t will be denoted D_t. The goods produced in period t using capital are $Y_t = (1+g)^{t-1} f(K_t/(1+g)^{t-1})$, where K_t is the period t capital stock, and f is an increasing, strictly concave function. This can be interpreted as a constant returns to scale production technology with a fixed factor that grows at the same rate g as other endowments. It will be assumed for simplicity that households are all equally endowed with the fixed factor.

Let us consider a stationary equilibrium in which at the end of each period, the households that had a high endowment in that period (type A in even periods, type B in odd periods) save by holding both government debt and capital, but the consumers who had a low endowment are liquidity constrained (i.e., would borrow if they could at the rate of interest received by savers). By a stationary equilibrium, I mean one in which $C_t^i/(1+g)^t$ equals $\bar{c}$ in every period t in which type i households have a high per capita endowment, and $\underline{c}$ in every period in which they have a low per capita endowment, in which $K_t/(1+g)^t = k$, $D_t/(1+g)^t = d$ and $T_t/(1+g)^t = \tau$ are likewise constants, and in which the real interest rate on government debt is a constant r. In such an equilibrium, $\bar{c}$, $\underline{c}$, k, d, τ, and r satisfy

$$v'(\bar{c})/v'(\underline{c}) = \beta(1+r) \tag{1}$$

$$v'(\underline{c})/v'(\bar{c}) \geq \beta(1+r) \tag{2}$$

$$f'(k) = 1+r \tag{3}$$

$$\bar{c} = e_1 + [f(k)-(1+r)k]/2 - \tau/2 - d - (1+g)k \tag{4}$$

$$\underline{c} = e_2 + [f(k)-(1+r)k]/2 - \tau/2 + d(1+r/1+g) + (1+r)k \tag{5}$$

$$\bar{c} + \underline{c} = e_1 + e_2 + f(k) - (1+g)k \tag{6}$$

Here the expression $[f(k)-(1+r)k]$ represents the competitive returns to the fixed factor of production. We wish to compare alternative stationary equilibria corresponding to different permanent levels d of outstanding debt per capita. This requires us to solve (1) and (3)–(6) for $\bar{c}$, $\underline{c}$, k, r, and τ, given d, checking that the solution is also consistent with (2). Since (1) implies (2) if $1+r \leq \beta^{-1}$, and is inconsistent with (2) otherwise, we can replace (2) by the requirement that the equilibrium real interest rate be below that upper bound.

These equilibrium conditions are identical to those of a neoclassical model of the kind considered by Peter Diamond (1965), with an appropriate identification of variables. Let us define $\tilde{c}_1 = \bar{c}$, $\tilde{c}_2 = (1+g)\underline{c}$, $\tilde{e}_1 = e_1$, $\tilde{e}_2 = (1+g)e_2$, and $u(\tilde{c}_1, \tilde{c}_2) = v(\tilde{c}_1) + \beta(1+g)v(\tilde{c}_2/1+g)$. Then equations (1) and (3)–(6) correspond to the conditions for a stationary equilibrium of an overlapping generations model in which each consumer lives for two consecutive periods, has endowment $\tilde{e}_1$ in the first period of life and $\tilde{e}_2$ in the second, has access to the same production technology as above, has an endowment of the fixed factor of production that is $(1+g)$ times as large in the second period of life as in the first, and chooses a life cycle consumption plan $(\tilde{c}_1, \tilde{c}_2)$ to maximize $u(\tilde{c}_1, \tilde{c}_2)$, and in which lump sum taxes in the second period are $1+g$ times as large as in the first.

As a result, the liquidity-constrained model predicts *exactly the same* effects of changes in the size of the government debt on real interest rates as does the neoclassical model, although the former model is perfectly consistent with the Barro view that finite lived consumers belong to infinite lived families linked by bequests that, as a result, act as though they were jointly maximizing a single infinite horizon objective function. In particular, if consumption is sufficiently substitutable over time, a higher stationary debt per capita d will be associated with a higher real interest rate r, and hence (because of (3)) with a lower stationary capital stock per capita. Nonetheless, the two models make different predictions about other kinds of policy experiments. In particular, the liquidity-constrained model predicts no necessary effect upon interest rates, saving, or capital accumulation of a change in the size of Social Security transfers, assuming that age is uncorrelated with whether a consumer is currently a saver or a dissaver, while the neoclassical model predicts that variations in the size of Social Security transfers have effects that are formally *equivalent* to those of a variation in the size of the public debt.[2]

Furthermore, the welfare consequences of variations in the size of the public debt are not the same in the two models. In the neoclassical model with perfect financial intermediation and lump sum taxes, equilibrium is Pareto optimal as long as $r \geq g$ (the case generally considered to be of empirical relevance, as discussed in Section III). And, comparing stationary equilibria satisfying this condition, all are Pareto optimal, but the stationary level of utility $u(\tilde{c}_1, \tilde{c}_2)$ is lower the higher is r, which is to say, the higher is d. The stationary level of utility is maximized by having a level of government debt only high enough to result in a rate of return $r = g$ (the Golden Rule case). But, in the model presented here, an efficient allocation of resources requires that $v'(\bar{c})/v'(\underline{c}) = v'(\underline{c})/v'(\bar{c})$ (i.e., that the liquidity constraints do not bind), which occurs only in an equilibrium with $1 + r = \beta^{-1}$. Thus efficiency requires that the real rate of interest be kept high enough, which, in the case of large enough intertemporal elasticity of substitution, requires that *the outstanding public debt per capita be maintained at a high enough level*. The level of real interest rates required for this condition to hold is higher than in the case of the neoclassical efficiency criterion, since (in order for the objective functions of the infinite lived households to be well defined) we must assume that $\beta^{-1} > 1 + g$. Furthermore, in the liquidity-constrained model, equilibrium is still inefficient (and a higher public debt is needed to achieve efficiency) *if it is still possible for an increased public debt to affect interest rates, saving, or investment*, since only when the liquidity constraints bind does Ricardian equivalence fail.

---

[2] In the neoclassical model, the relation that exists between d and variables such as r and k depends upon the distribution of net tax collections between young and old, which is why Kotlikoff argues that the deficit as such is irrelevant. In the liquidity-constrained model, the relation that exists between d and r and k similarly depends upon the distribution of net tax collections between liquidity-constrained and unconstrained households, but, in this case, there is much greater reason to set aside the possibility of changes in that ratio when considering optimal fiscal policy (the liquidity constraints themselves may exist because of difficulty in observing differences in individual households' circumstances), and, in any event, the additional dimension of fiscal policy represented by that ratio has nothing to do with Social Security provisions.

This result is not really a novel one. For the model is exactly the one in which Truman Bewley (1980) explains Friedman's doctrine regarding the "optimum quantity of money." Bewley postulates that money is held because it supplies liquidity in exactly the sense discussed here—it allows consumers to smooth consumption in the face of endowment fluctuations and an assumed inability to borrow against future endowment income. He compares alternative stationary equilibria with different constant rates of money growth and inflation, and argues that a policy that results in a higher level of real money balances and a higher real return to holding money improves efficiency. Specifically, he shows that a Pareto optimal allocation of resources occurs only when the real return on money is made as high as households' rate of time preference. This is just my result regarding the return on government debt. Note that my model has said nothing about nominal values; the above conclusions are independent of the rates of inflation and hence of nominal interest rates in the alternative stationary equilibria, and, in particular, continue to hold in the case of government debt bearing a zero nominal interest rate.

According to a common view, the Bewley-Friedman argument is relevant for money, but not for other government paper. However, Treasury securities (especially those with short maturities) also seem to enjoy a liquidity premium. This is shown by the fact that the yield on Treasury bills is lower than that on other assets, such as equities, to an extent that cannot plausibly be accounted for simply as a consequence of the assets' different degree of riskiness. But the existence of a liquidity premium implies perfect financial intermediation, and a fiscal policy that lowers this premium by bringing the rate of return on Treasury bills into line with those on other securities should increase efficiency.[3]

One disadvantage of the above model as an illustration of this view is that government debt and capital must earn the same rate of return, so that the liquidity services provided by government debt do not show up as a liquidity premium on government debt relative to other assets, and the increase in efficiency associated with a higher-debt policy is not reflected in the reduction of such a premium. However, as the following example shows, such effects can easily occur in a liquidity-constrained economy.

II. Public Debt May "Crowd In" Investment

I have presented an example above in which the effects of government debt on national saving and investment are exactly

[3] This conclusion is subject, of course, to the same sorts of second-best considerations that have been pointed out as qualifications to the Friedman argument. (See my forthcoming paper for a survey.)

the same as in the neoclassical model, to show that even if those predictions of the neoclassical model were found to be correct, the welfare conclusions of the model could be wrong. But, in fact, a higher public debt does not imply a lower capital stock under such general conditions in the case of economies with imperfect financial intermediation as in the neoclassical model. This is another significant difference between the two models. It is important to realize that acceptance of the view that the high real interest rates of the 1980s are largely a consequence of the change in U.S. fiscal policy does not require one to also believe that the higher government deficits have crowded out investment.[4] The following variant of the previous model shows how a higher public debt can actually bring about higher levels of national saving and investment, by reducing the extent to which people with access to productive investment opportunities are liquidity constrained.

Instead of assuming that consumers have access to the production technology in every period, let us suppose that a given household has access to it only in certain periods. The idea (represented here in an extreme form) is that particularly attractive investment opportunities come along for a given economic unit only at certain times, so that it has more use for funds on those occasions than at other times. An important function of liquid assets, in an economy without frictionless loan markets, is to allow such entities to concentrate their spending more in the periods when they have especially good opportunities. In order to focus upon this function, we can ignore altogether the need to smooth endowment fluctuations, by assuming a constant endowment stream.

Let us again suppose that there are two types of infinite lived households, with the same preferences and endowments as before, except that now $e_1 = e_2 = e$. The two types now differ instead in the times at which they have access to the production technology. Type A households can invest in physical capital in odd periods, and use that capital to produce in the following even periods; type *B* households have the opportunity to invest in even periods and to produce in the following odd periods. The production technology is the same as in the previous section. Again, the endowment of the fixed factor is equally divided across the two household types. Finally, let us assume again that private borrowing is impossible, but that all households are able to save by holding government debt. Taxes are again lump sum and equally distributed across household types.

Let us consider a stationary equilibrium in which all consumers are liquidity constrained in the periods in which they have investment opportunities (and as a result, hold no government debt in those periods), but not in the other periods (in which they save by holding government debt). Again, let $\bar{c}$ denote consumption per family member when the household is not liquidity constrained, and $\underline{c}$ consumption per family member when it is. Let d denote government debt held at the end of the period, and k denote the capital stock brought into the period, per non-liquidity-constrained family member, and let $\tau/2$ denote the taxes collected per family member each period. Then in such a stationary equilibrium, $\bar{c}$, $\underline{c}$, d, k, τ, and r must satisfy

$$v'(\bar{c})/v'(\underline{c}) = \beta(1+r) \tag{7}$$

$$v'(\underline{c})/v'(\bar{c}) = \beta f'(k) \tag{8}$$

$$f'(k) \geq 1+r \tag{9}$$

$$\bar{c} = e_1 + [f(k)-(1+r)k]/2 - d + (1+r)k - \tau/2 \tag{10}$$

$$\underline{c} = e_2 + [f(k)-(1+r)k]/2 + d(1+r/1+g) - (1+g)k - \tau/2 \tag{11}$$

$$\bar{c} + \underline{c} = 2e + f(k) - (1+g)k \tag{12}$$

---

[4] The other major industrial countries, that have on the whole pursued tighter fiscal policies, have suffered even greater declines in rates of private saving and domestic rates of investment (Barry Bosworth, 1990). He suggests that the reduced rates of saving and investment have resulted mainly from the slowdown in income and productivity growth since the mid-1970s.

Let us consider how the solution for $\bar{c}$, $\underline{c}$, k, τ, and r varies as the value of d is varied.

The most important difference between this model and that of the previous section is that (7) and (8) imply

$$f'(k) = \beta^{-2}(1+r)^{-1} \tag{13}$$

so that the steady-state capital stock varies directly, rather than inversely, with the real return on government debt, which here is not equal to the return on capital. In fact, the spread between the return on capital and that on government debt is given by $f'(k)-(1+r)$, which by (13) is a decreasing function of r. As before, if consumption is sufficiently substitutable between periods, stationary equilibria with higher values of d (in the range where the liquidity constraints bind) will be associated with higher levels of the real interest rate r. But, now this implies a *higher* stationary capital stock per capita. A higher government debt can actually increase the steady-state capital stock, by improving the efficiency with which investment can be financed as a result of easing households' problem of the illiquidity of their claims to future endowment income.[5] Furthermore, in the case described, the spread between the return on capital and on government debt will be a decreasing function of the level of debt. Thus the increase in private sector liquidity due to an increase in public debt shows up in this case as a reduction of the spread between these two rates of return.

As in the previous section, in this economy Pareto optimality requires that $v'(\bar{c})/v'(\underline{c}) = v'(\underline{c})/v'(\bar{c})$, which again requires that $1+r=\beta^{-1}$. Again, keeping the real return on government debt high enough (in this case, high enough to eliminate the spread between the returns on the two assets) requires that the public debt be maintained at a high enough level.

III. Higher Public Debt Need Not Imply Higher Taxes

An obvious qualification to the results of the previous two sections concerns the assumption of lump sum taxation. Equations (4)–(6), or equivalently (10)–(12), imply the relation $\tau = (r-g/1+g)d$. If $r>g$, this implies that a higher-debt stationary equilibrium must also have a higher level of tax collections. If, as assumed thus far, these taxes are lump sum, the size of tax collections has no effect upon welfare comparisons. But, it is more realistic to assume that the taxes must be raised in ways that will distort the allocation of resources. This would be a source of lower welfare in high-debt stationary equilibria, and would have to be weighed against the liquidity effects discussed above.

While this argument could well mean that it is not optimal to increase the size of the public debt to the point of "satiation in liquidity" (i.e., the point where liquidity constraints cease to bind), it remains likely, even when $r>g$, that some positive permanent level of public debt per capita will be optimal. This is in contrast to it being desirable to reduce the debt as much as possible. But, it is also important to realize that the extent to which a permanent increase in the ratio of public debt to national income requires an increase in future taxes need not be very great. Indeed, there may not have to be any increase in taxes at all.

Higher d implies higher τ only if $r>g$. Andrew Abel et al. (1989) argue that this is the empirically relevant case, in the context of a standard neoclassical model with efficient financial intermediation. In a stationary equilibrium with $r<g$, gross profits $k_t f'(k_t)$ should in each period be less than gross investment $(1+g)k_{t+1}$; but, in the U.S economy, gross profits are always about 25–28 percent of GNP while gross investment is only about 14–18 percent of GNP. But this evidence actually only shows that $f'(k)>1+g$. In an economy with efficient financial intermediation, this would imply that $r>g$, but in a liquidity-constrained economy like that discussed in the previous section, it does not. Dynamic efficiency of the intertemporal production plan, established by Abel et al., does not rule out the possibility that an increase in the public debt can be simply rolled over forever without taxes ever having to be increased. Indeed, this appears to be a realistic possibility in the case of the U.S. economy. For the average real return on Treasury bills over the postwar period has been close to zero, while the average growth rate of real GNP has been over 3 percent per year. If this is so, the liquidity effects described in the previous two sections suggest that welfare could be increased by a permanent increase in the level of the public debt.

---

[5] Hence Eisner's result suggesting that increases in the public debt tend to be associated with increases, rather than decreases, in investment is consistent with a fully "classical" theory of aggregate supply determination.

REFERENCES

Abel, Andrew B. et al., "Assessing Dynamic Efficiency: Theory and Evidence," *Review of Economic Studies*, January 1989, *56*, 1–19.

Barro, Robert J., "The Ricardian Approach to Budget Deficits," *Journal of Economic Perspectives*, Spring 1989, *3*, 37–54.

Bernheim, B. Douglas, "A Neoclassical Perspective on Budget Deficits," *Journal of Economic Perspectives*, Spring 1989, *3*, 55–72.

Bewley, Truman F., "The Optimum Quantity of Money," in J. H. Kareken and N. Wallace, eds., *Models of Monetary Economies*, Minneapolis: FRBank of Minneapolis, 1980.

Bosworth, Barry P., "International Differences in Saving," *American Economic Review Proceedings*, May 1990, *80*, 377–81.

Diamond, Peter A., "National Debt and Neoclassical Economic Growth," *American Economic Review*, December 1965, *55*, 1125–50.

Diaz-Gimenes, Javier, and Prescott, Edward C., "Asset Returns in Computable General Equilibrium Heterogeneous Agent Economies," unpublished, Research Department, FRBank of Minneapolis, December 1989.

Eisner, Robert, "Budget Deficits: Rhetoric and Reality," *Journal of Economic Perspectives*, Spring 1989, *3*, 73–93.

Huggett, Mark, "Risk-Free Rates in Heterogeneous-Agent Economies," unpublished, University of Minnesota, October 1989.

Kotlikoff, Laurence J., "From Deficit Delusion to the Fiscal Balance Rule: Looking for an Economically Meaningful Way to Assess Fiscal Policy," unpublished, Boston University, December 1988.

Poterba, James M., and Summers, Lawrence H., "Finite Lifetimes and the Savings Effects of Budget Deficits," *Journal of Monetary Economics*, September 1987, *20*, 369–91.

Woodford, Michael, "The Optimum Quantity of Money," in B. Friedman and F. H. Hahn, eds., *Handbook of Monetary Economics*, Amsterdam: North-Holland, forthcoming.

DEREGULATED AIRLINE MARKETS[†]

The Dynamics of Airline Pricing and Competition

By STEVEN A. MORRISON AND CLIFFORD WINSTON[*]

Like the prices of many goods and services, air fares exhibit cyclical fluctuations. Unlike other price fluctuations, air fare cycles spark debate about industry policy. The rise in fares during the most recent cycle prompted a congressional investigation of competition at hub airports, provoked the Secretary of Transportation to stress that proposed airline mergers would be evaluated much more carefully, and encouraged Congress and the Secretary to indicate their opposition to leveraged buyouts that might lead to bankruptcy. These official reactions have been accompanied by popular concern with airline deregulation and an interest in some form of reregulation.

According to our investigation, during the past decade of deregulation fares were consistently lower than they would have been if the industry were still regulated. To address concerns about recent fare increases, we examined whether market forces are operating as freely as possible in fare determination. Among our findings are that fares fall with increased competition, but this effect is limited at slot-controlled airports, and that an airline's entry and exit behavior is significantly influenced by its own (but not its competitors') networks. We discuss the implications of these findings for public policy at the conclusion of the paper.

I. Fares under Deregulation

The trend in U.S. domestic air fares since deregulation in 1978 is shown in Figure 1. Yields (revenue per passenger mile) were calculated from the Department of Transportation *Ticket Dollar Value Origin and Destination Survey* (Data Bank 1A). Despite significant fluctuations, the nominal yield at the beginning of 1987 was lower than the nominal yield during 1980. Fares climbed substantially during 1987 and 1988; the increase in 1988 alone was 20 percent. But during 1989 fares turned down.

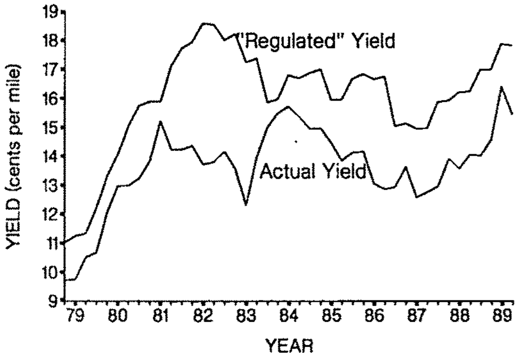


FIGURE 1. ACTUAL VS. "REGULATED" YIELD

Skeptics of deregulation see rising fares as evidence that deregulated competition has failed to replace the discipline that the Civil Aeronautics Board (CAB) once placed on carriers. We used the Standard Industry Fare Level (SIFL), used during regulation by the CAB to set fares (and still calculated by the Department of Transportation for other purposes), to predict what fares would have been if they were still regulated. It was necessary to recalculate the SIFL using only those airlines that were included in its calculation under regulation, and to adjust the costs that enter the SIFL so that they did not reflect changes in productivity growth and in factor prices that are attributable to deregulation.[1]

[†] *Discussants*: Nancy Rose, MIT; Elizabeth Bailey, Carnegie Mellon University; Ariel Pakes, Yale University.

[*] The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, DC 20036 and Northeastern University, and The Brookings Institution, respectively.

[1] Details underlying this calculation are available from the authors.

Yields during the past decade if airlines had been regulated are also shown in Figure 1.[2] Although the gap has narrowed in the last few years, regulated yields always exceed actual yields. On average, deregulated fares are lower than regulated fares by 18 percent, amounting to an average annual saving to travelers of roughly $6 billion (1988 dollars). Of course, yields disguise considerable variance in deregulated fares; for some travelers and on some routes fares are significantly greater than regulated fares would have been. For example, because fares on long-distance routes were deliberately set very high under regulation (and short routes had low fares) deregulation reduced fares for trips that exceeded 900 miles (one-way) and increased them for trips less than 900 miles.

II. Competition and Fares

Critics of deregulation claim that fares have increased, especially during the past few years, because competition has eroded. Although the number of effective competitors at the national level has fallen from 8.7 in 1978 to 7.7 in 1988 (with a peak of 11.2 in 1985), the number of effective competitors at the route level, where competition actually takes place, has risen from 1.52 to 1.90 during this period (with a peak of 1.96 in 1986).[3] Furthermore, the percentage of travelers flying on carriers with 90 percent or greater market share has fallen from 28 percent in 1978 to 17 percent in 1988; the percentage of travelers flying on carriers with 20 percent or less market share has risen from 7 to 17 percent during the same period.

Like fares, the number of effective competitors has changed during the decade. To what extent does competition affect fares, and should public policy simply focus on increasing the number of competitors in a market? Or are there other factors underlying the relation between fares and competition?

TABLE 1—FARE REGRESSION

Variable	Coefficient	Standard Error
Distance (miles)	.501	.003
Number of effective competitors on routes without slot airports		
1978–81	−.037	.010
1982–88	−.119	.006
Number of effective competitors on routes with slot airports		
1978–81	.006	.013
1982–88	−.035	.008
Minimum number of effective competitors at a route's endpoints		
1978–81	−.015	.009
1982–88	−.201	.005
Potential carriers		
1978–81	$-.55E{-}02$	$.06E{-}02$
1982–88	$-.14E{-}02$	$.04E{-}02$
$n = 18{,}573$; $R^2 = .99$		

Note: All variables are natural logarithms.

Several writers have addressed these questions by estimating cross-section regressions of fares as a function of potential and actual competitors, route characteristics, and airport characteristics. We employed this specification to investigate whether the relation between fares and competition has remained stable over time using a pooled cross-section time-series sample of route activity from 1978:4 to 1988:4.[4] We included separate constants for each year, but to save space, we did not include them in the results presented in Table 1.

Based on regressions not shown here, the impact of the number of effective airlines in a market on fares has remained stable following the transition to full deregulation of entry in 1982 and was not lower on routes involving a carrier's hub. For routes that do not have a slot-controlled airport, the long-run elasticity in Table 1 is $-.12$, which is consistent with previous cross-section evidence. For routes involving a slot-controlled airport, the long-run elasticity falls to $-.04$. An explanation for this finding is that fares on slot-controlled routes reflect the opportunity cost of the slot (that depends on total airport demand) and not just on competition on any single route. We also found as an airport (endpoint) became more concentrated, fares rose. This effect has grown during deregulation, possibly because of frequent flier programs, while the impact of potential competition in reducing fares has fallen.

---

[2] The initial gap between "regulated" yield and actual yield reflects deregulation induced productivity changes and fare cuts that took effect before statutory deregulation in October 1978. Comparison of regulated yield and *fully* deregulated yield should be made beginning in 1983.

[3] Effective competitors refers to the number of "equivalent equal size competitors" calculated by inverting the appropriate Herfindahl index.

[4] It has been argued on theoretical and empirical grounds that competition can be treated as exogenous in this specification, but there is some resistance to this assumption. Our entry model below suggests that to the extent competition is influenced by fares, the appropriate variable is the *relative* fare and not the actual fare in the market. Data for this regression are from Data Bank 1A.

III. Entry and Exit

We have corroborated that an increase in airline competition does indeed lower fares. How does competition arise? This question has become especially important because during 1987:1–1989:1, industry fares increased by 30 percent while factor prices increased by 13 percent, prompting concern about the existence of entry barriers and the adequacy of competition, particularly on routes with airports dominated by one carrier.

We analyzed the determinants of a carrier's decision to provide direct service, identifying the importance both of its own, and of other carriers', activity at the airports in the market. Using the Service Segment data base, we constructed a sample of quarterly entry decisions for 1979:1–1988:4 by 13 airlines formerly classified as trunk and local service carriers. For a given quarter, carriers were given the opportunity to enter any sampled market receiving airline service that they did not serve in the previous quarter. The influence of a carrier's network on its entry decision was captured by the carrier's maximum share of total enplanements at the origin and destination airports during the previous quarter (*OWNNET*). The influence of its competitors' networks on its entry decisions was captured by the maximum share of any other carrier's total enplanements at the origin and destination airports (*COMPNET*). The specification also accounts for the influence of fares on entry, as defined by the yield on the route in the previous quarter divided by the yield for routes of the same distance in the previous quarter *RELFARE*).[5] The effect of potential traffic volume on entry is accounted for by the product of the origin and destination populations (*POP*). Finally, we included the number of slot-constrained airports (*SLOTS*) on the route. Data for the explanatory variables were from the *Origin-Destination* Data Bank 1A.

The expected signs of all the variables except *RELFARE* are straightforward: *OWNNET* and *POP* should have a positive effect on the likelihood of entry; *COMPNET* and *SLOTS* should have a negative effect on the likelihood of entry. A natural expectation of the effect of *RELFARE* is that it will be positive—relatively higher fares attract entry. However, if fares are set to deter entry, *RELFARE* should have a statistically insignificant effect. Alternatively, high fares could signal entry barriers (say, frequent flier programs) or indicate that incumbent carriers will respond aggressively to entry; thus *RELFARE* would have a negative effect on entry. Relatively higher fares could also signal relatively higher costs.

We estimated probit entry models for each of our carriers during partial deregulation (1979:1–1982:4) and full deregulation (1983:1–1988:4). To save space, we present our findings for American Airlines in Table 2 and describe the findings for other carriers in conjunction with these results. American's activity at origin and destination airports (*OWNNET*) had a significant impact on its entry decisions. Indeed, this effect was statistically significant for all carriers. In some cases its magnitude grew during the transition to full deregulation, and in others, like American's, it fell. Other carriers' activity at origin and destination airports

---

[5] *RELFARE* is set by *all* carriers in a market before each carrier (not serving the market) makes its entry decision. *RELFARE* is thus unlikely to be caused by an entry decision.

TABLE 2—PROBIT ENTRY MODEL PARAMETER ESTIMATES: AMERICAN AIRLINES

	1979–82	1983–88
Constant	−1.18	−1.30
	(.136)	(.104)
OWNNET	1.90	1.05
	(.252)	(.221)
COMPNET	−1.77	.011
	(.207)	(.141)
RELFARE	−.032	−.252
	(.113)	(.083)
POP	.242*E*-07	.335*E*-07
	(.380*E*-08)	(.364*E*-08)
SLOTS	−.116	−.367
	(.073)	(.075)
No. of observations	7217	5366
Log likelihood at zero	−5002	−3703
Log likelihood at convergence	−1246	−1607

Note: Standard errors are shown in parentheses.

(*COMPNET*) discouraged entry by American during partial deregulation, but by full deregulation American's entry decisions were unaffected by this activity. This finding persisted for the vast majority of carriers. It appears that carriers' networks have become sufficiently developed that entry decisions are dictated by their own strength in particular markets without consideration of the activity of other carriers.

During full deregulation, most carriers' entry decisions, including American's, were negatively affected by *RELFARE*, suggesting that fares signal entry barriers, relatively high costs, or incumbents' response to entry, or all three. For most carriers the magnitude of *RELFARE* increased during the transition from partial to full deregulation, when entry barriers such as frequent flyer programs were firmly established. However, this does not imply that high fare routes are immune from entry. In conjunction with the *OWNNET* finding, a carrier that wishes to enter a high-fare route needs a strong network: a fare 10 percent above average is offset for American Airlines by a 2.4 point larger share of traffic at the origin or destination. Finally, as expected, *POP* and *SLOTS* have a positive and negative effect, respectively, on entry.

As an extension and a check of the internal consistency of our findings, we estimated a probit model of exit behavior. The signs for *OWNNET* and *COMPNET* were reversed, as expected, and *COMPNET*'s influence generally declined under full deregulation. *OWNNET*'s effect was always statistically significant and far greater than its effect on entry. Its elasticity for entry was usually close to .1, but its elasticity for exit was typically around −.5 and as high as −1.0. This suggests that one of the most important effects of a well-developed airline network is that it preserves competition on routes that are served. *RELFARE* had a positive effect on exit, perhaps indicating that carriers that leave markets with relatively high fares face promotional disadvantages or anticipate aggressive behavior by the other carriers.

IV. Policy Implications

One immediate policy implication of our analysis is that although CAB-style fare regulation could eliminate the spread in fares for similar trips, it would not lower the level of fares. Even though yields did rise dramatically during 1987–88, fares would have been greater had they been regulated as before. Although in principle a more "enlightened" form of fare regulation could reduce the spread of fares without increasing their level, it is difficult to believe such benign regulation could occur in practice.

Public policy should focus on enhancing the effect of competition on fares and on increasing the number of competitors in markets. Because slots limit the effect of competition on fares and the number of competitors, they should be eliminated and replaced by congestion-based takeoff and landing fees. Congestion pricing would reduce travel delays efficiently (see our 1989 paper), and could enhance competition.

We found that fares are higher on routes with greater carrier concentration at airports. This problem is believed to be most serious when one of the airports is a carrier's hub. A common explanation is that hubs pose entry barriers and therefore limit competition. We have found that hubs do not limit the *effect* of competition on fares, and that they no longer explicitly affect carriers' entry decisions. But because airlines initially developed their networks from their historical "hubs," they have been slow, especially during partial deregulation, to expand into routes that are connected to competitors' hubs. But as carriers continue to expand their networks, they should come into direct competition more often on routes where one of the airports is a competitor's hub. Further, because they will be entering these routes from a position of strength, they will be less likely to exit them.

To be sure, hubs may exacerbate barriers to entry created by frequent flyer programs, an effect that may have been captured in our analysis by the negative effect of fares on entry. But an appropriate (although difficult) policy response, which could benefit air travelers (see our 1989 paper), is to tax or eliminate frequent flyer programs.

Because hubs have a beneficial effect on a carrier's entry and exit behavior, and because they enable carriers to provide better service (see our 1986 study), it would be inadvisable to deter the formation of new hubs or dismantle existing ones. The higher fares associated with hubs should erode as carriers expand their networks. And if future entry by foreign airlines (cabotage) is to help lower fares, these carriers must be allowed to form hubs from which to enter domestic routes. Ironically, abolishing hubs in the hope of promoting competition may achieve the opposite effect.

REFERENCES

Morrison, Steven A. and Winston, Clifford, *The Economic Effects of Airline Deregulation*, Washington: Brookings Institution, 1986.

______ **and** ______, "Enhancing the Performance of the Deregulated Air Transportation System," *Brookings Papers on Economic Activity: Microeconomics*, 1:1989, 61–112.

Airport Presence as Product Differentiation

By STEVEN T. BERRY*

There is now widespread recognition that an airline's operation at a given airport greatly affects its competitive position on routes flown out of that airport (see M. Levine, 1987; S. Borenstein, 1989; S. Morrison and C. Winston, 1989; and myself, 1989, among many others.) Airlines typically defend any competitive advantage as stemming from the lower costs and better service that are said to be generated by hub-and-spoke route systems. Airline critics typically respond that airlines gain market power by dominating individual airports.

Both views are at least plausible. Hub-and-spoke transportation networks reduce the number of round-trips necessary to carry a given number of passengers on a given set of itineraries, while increasing the number of passenger miles flown. If there are sufficient economies of scale in plane size, then the advantages of hubbing can overcome the disadvantage in passenger miles, resulting in lower total costs. By pooling passengers with different ultimate destinations, a hubbed system can also offer more frequent flights than would be economically feasible under a nonstop system.

It also appears, however, that airlines gain other advantages from a large presence at an airport. Incumbent airlines are the major source of financing for many airports and therefore gain a large degree of bureaucratic control over airport operations. This control may enable them to block the entry or expansion of rivals. Airlines with a large presence in a given city also gain advantages from frequent flyer plans and nonlinear travel agent commission schedules (see, again, Levine and others). If the bureaucratic and marketing advantages of airport presence are sufficient to prevent most attempts at entry, then incumbents may gain the ability to exercise traditional "market power" by restricting output and driving up prices.

This paper argues that both simple cost-reducing and naive market power stories are inappropriate for the airline industry. I present a model in which consumers are willing to pay a premium for the services of the dominant airline; this premium may be related to a number of factors, including flight frequency, frequent flier miles, and travel agent commission overrides. This model has the advantage of treating oligopoly product differentiation in an explicit way, of treating price as an endogenous variable, and of allowing for airport presence to affect both costs and demand.

I. Prior Empirical Evidence

Empirical evidence for the importance of serving a large number of routes out of a given city comes from studies of pricing, demand, and entry. These studies have typically focused on showing that there are important effects of airport dominance, without embedding the empirical work in an economic context that might distinguish demand-side from cost-side advantages.

Borenstein (1989) has shown that airlines charge higher prices on flights involving their hub airports. Furthermore, prices are higher on flights from hub airports as opposed to flights travelling to hub airports. This is consistent, for example, with important frequent flyer effects.

Morrison and Winston use a traditional logit framework to analyze demand for airline services and find that, holding price constant, more customers choose airlines who have large operations out of the origin city. This analysis has a number of serious flaws, including the treatment of price as an exogenous variable and treating the aggregate demand for flights (across all firms) as perfectly inelastic. However, as we will see below, their basic descriptive result reappears in a somewhat more sophisticated demand model.

---

*Yale University, New Haven, CT 06520. Support from an Eno Foundation grant to the Department of Economics is gratefully acknowledged. Maureen Germani provided excellent assistance with the data. I thank Ariel Pakes and Mike Whinston for helpful conversations on this topic.

My earlier paper studied airline profitability in city-pairs by treating entry decisions as an indication of underlying profitability. This work finds that entry decisions are highly correlated with airport presence. The importance of airport presence in entry decisions (and by inference, in profits) is also supported in this work by the estimation of a game-theoretic model of entry (for results in the same spirit, see also P. Reiss and P. Spiller, 1987).

My earlier paper made no attempt to discover the source of the profitability advantage that is associated with dominant airlines, but the studies of pricing and demand provide at least indirect support for the importance of product differentiation. If hubbing decreases costs, without any other effect, every standard market model of which I am aware predicts that airline prices should fall as hubbing increases. By contrast, a traditional model of "market power" predicts high prices with low output. In fact, the papers cited above indicate that airlines that dominate the traffic out of a given city appear to charge higher prices than other firms on the same routes while serving a larger number of passengers; this is at least consistent with explanations based on product differentiation.

II. Prices and Quantities

This section discusses the data and presents some reduced-form results that are consistent with the work described above, and with the structural model presented in the next section. First, I consider a (linear) reduced-form equation for price and quantity in a traditional "supply and demand" framework. Exogenous factors that increase demand but do not affect costs will have a positive coefficient in both the quantity and price equations. Exogenous factors that increase costs, but do not affect demand will have a positive coefficient in the price equation and a negative coefficient in the quantity equation. Without more structure, we can make no reduced-form predictions for factors that affect both cost and demand.

TABLE 1—OLS ESTIMATES OF PRICE AND QUANTITY[a]

	Equations	
Coefficient	Price	Passengers
Constant	1.81	−2.47
	(1.35)	(0.341)
Population	0.118	4.92
	(.0895)	(0.227)
Distance	1.05	−0.559
	(0.0631)	(0.160)
Distance^2	−0.0633	0.0440
	(0.0101)	(0.0255)
Routes	0.146	0.649
(origin)	(0.0232)	(0.588)
Routes	−0.226	0.701
(dest.)	(0.0218)	(0.0553)
Hist. Cost	16.3	−14.7
	(2.06)	(5.22)
New Line	−0.714	2.27
	(0.0889)	(0.225)

[a] Standard errors are shown in parentheses.

Table 1 presents the results from ordinary least squares (OLS) estimation of such reduced-form equations for the airline industry in 1984. The sample is from the *Origin and Destination Survey* (a 10 percent sample) of air passenger traffic and consists of 300 city-pair markets formed from all the possible pairs connecting 25 of the largest U.S. cities. Following Borenstein (1990), travel from point A to point B is taken as a different market from travel from point B to point A. This gives 600 markets in the sample. In Table 1, an observation is an airline/market combination; the next section will place airlines into markets. The sample consists of round-trip itineraries on a single airline, and includes both nonstop and multiple stop itineraries.

The dependent variables are the average price charged by the airline on the given route (in hundreds of dollars) and the number of passengers flown by the airline (in hundreds of sampled tickets). Population (entered as the product of the two cities population, in tens of millions) is assumed to affect demand but not cost. Distance between the two cities (in thousands of miles) is assumed to affect both demand and airline costs.

Other potential cost variables raise questions of potential endogeneity. For example, profit-sharing models of wage determination raise important questions about the exogeneity of current wages. Exogenous cost factors are modeled here as depending on the history of the airline; the airlines of the regulated period are assumed to have inherited higher costs from that era of service competition and fixed prices. Specifically, the regressions include a variable which indexes historical firm costs for the major trunk carriers of the regulatory era. This variable (*Hist. Cost* in the tables), is taken from D. Caves et al. (1984) and consists of the estimated firm-specific effect (net of wages and other measured inputs) from an estimated industry cost function covering the years 1970–82. Another variable (New Line in the tables) is equal to one for firms other than the old trunk lines; I expect cost to be lower for these firms.

Airport presence, measured crudely by the number of top 50 cities served by an airline out of a given city, is allowed to affect both cost and demand. Since several of the product differentiation stories apply to presence at the origin city, the number of routes served variable is entered separately for the origin and destination cities.

Results are as expected, if somewhat difficult to interpret in this reduced-form setting. The historical cost variable increases prices and decreases demand. The effect of airport presence varies between origin and destination. The signs of variables that affect both cost and demand are mixed. Interestingly, population affects the number of passengers carried, but not the price charged; this is a direct implication of the explicit product differentiation model estimated in the next section.

III. A Structural Model of Product Differentiation

The reduced-form analysis above cannot decompose the effect of airport presence into demand and cost effects. If airline markets were perfectly competitive, then we could estimate "structural" cost and demand equations separately, instrumenting the cost side with omitted demand variables (here, *Population*) and instrumenting the demand side with omitted cost variables (here, the two variables measuring historical cost.) However, it is not so easy in an oligopolistic differentiated products framework. The price of every firm in a city-pair should depend in some way on the prices of all other firms. In the reduced form, the price of a given firm will depend on the characteristics of all the firms in the market. Since there can be more than a dozen firms serving a city-pair market, the number of parameters for arbitrary demand specifications becomes unmanageable. It is necessary to impose more structure on the model to reduce the complexity of the problem.

Several approaches to structural modeling of demand are possible. Discrete choice models of consumer behavior (such as logit models) have the advantage that they allow for an explicit relationship between product characteristics and consumer utility which is parameterized in a parsimonous fashion. By allowing for an unobservable component of a consumers preference for a good, they allow consumers to disagree on the identity of the most desirable product.

However, discrete choice models have several disadvantages in the present context. When the number of consumer choices (here, airlines) is more than two or three, the exact calculation of market shares is difficult for most specifications other than logit. This problem is largely solved by the recently introduced techniques of simulation estimators (see A. Pakes and D. Pollard, 1989, and D. McFadden, 1989).

Also, discrete choice analysis of a single market typically involves unobservables only at the level of the consumer. Any unobserved product characteristics are assumed to be captured by the intercept term of product utility. In this single-market framework, aggregate market shares should be predicted with near perfection because the individual errors will tend to average out over a large number of observations. When we are dealing with a large number of markets, it is unreasonable to believe that the econometri-

TABLE 2—THE PRODUCT DIFFERENTIATION MODEL

	Estimate	Standard Error
Market Size (M)		
Constant	2.24	(2.86)
Population	45.8	(11.4)
Distance	4.06	(2.64)
Distance^2	0.596	(0.541)
Utility (μ, α)		
Constant	−1.85	(2.20)
Routes (origin)	0.665	(0.0191)
Routes (dest.)	0.315	(0.0229)
Price ($-\alpha$)	−1.32	(−2.94)
Marginal Cost		
Constant	2.59	(0.237)
Distance	0.715	(0.0922)
Distance^2	−0.0237	(0.0156)
Hist. Cost	2.07	(3.29)
New Line	−0.359	(0.252)
Routes (origin)	−0.0750	(0.0325)
Routes (dest.)	−0.269	(0.0399)
Std Deviation of Unobservables		
Demand (ξ)	1.85	(0.108)
Cost (ν)	0.001	(0.002)

cian observes all the variation in product characteristics across markets, so we would typically believe that there are market-specific unobservables. Furthermore, this market-specific demand error will be correlated with price in equilibrium so it is not appropriate to treat price as an exogenous variable.

Finally, in an oligopoly framework, the first-order condition for profit-maximizing prices will include the derivative of the demand function with respect to own-price. The calculation of predicted price requires precise knowledge of this derivative, which is difficult to calculate analytically in most discrete-choice models. Even when this derivative is easy to calculate (as in logit), the resulting supply relationship is not easy to solve for equilibrium prices; indeed, there is no guarantee that equilibrium prices exist in the general case.

Table 2 presents preliminary results from the estimation of a simple product differentiation model that takes these problems into account. In this model, the mean valuation of an airline in a given city-pair depends on observed characteristics, but individual consumer valuations of the firm have a distribution about that mean. Firm characteristics depend on observable market presence and on an unobserved term. The resulting demand curve is not easy to calculate and a simulation estimator must be used in estimation. However, unlike logit or probit, the equation for equilibrium prices takes a particularly easy form that eases the computational burden considerably.

I give only the assumptions and the results of the model here (an appendix is available on request from the author). Consumers choose one unit of the product that gives the highest utility. In a given market, utility for consumer i of choosing firm j is $u_{ij} = v_{ij} - \alpha p_j$, where v_{ij} is assumed to be exponential with mean μ_j and $\mu_j = \exp(x\beta + \xi_j)$. The parameter α is the marginal utility of income while the random variable ξ_j is a market-specific unobservable, assumed to be normally distributed with zero mean conditional on the exogenous data. The "outside" value of not consuming is also assumed to have an exponential distribution (with mean normalized to one), so that an overall increase in prices will drive consumers from the market.

The exponential assumption is arbitrary, but a skewed distribution may be reasonable —while most consumers place a very low value on the product, a few place a very high value. More practically, the exponential discrete choice model gives a simple form for equilibrium prices: $p_j = mc_j + \mu_j/\alpha$, where mc_j is marginal cost (assumed to be constant in output quantity for simplicity) and (μ_j, α) are the demand parameters described above. The derivation of this result is similar to that of constant elasticity demand curves in the monopolistic competition literature (see, M. Sattinger, 1984). Marginal cost mc_j is parameterized as $w\gamma + \nu_j$, where ν_j is a normal random variable with zero mean conditional on the exogenous data and γ is a parameter.

The simple reduced form for price in this model eases computational burdens, but it is restrictive. In particular, equilibrium price is found to be unaffected by the characteristics of other firms; thus, for example, entry will reduce the market shares of existing firms but not their prices.

The model as outlined predicts market shares and equilibrium prices. However, not observed is the number of consumers who choose not to fly at all. We must therefore model the total number of potential consumers; demand for a firm is then just its market share multiplied by the market size, denoted M. Market size is modeled in Table 2 as a linear function of population and distance: $M = z\delta$, where δ is a parameter vector to be estimated and z contains the market size variables.

The results in Table 2 are based on the same data as the reduced-form results in Table 1. Parameter estimates are obtained from a simulation estimator. In this technique, a predicted price is calculated from the equilibrium price equation and this predicted price is used to simulate predicted market shares based on the assumed distributions of ξ and ν. The parameters are then chosen to match the observed and predicted prices and quantities as closely as possible. (My earlier paper gives another example of simulation estimators applied to a game-theoretic problem in the airline industry).

The results in Table 2 appear to be sensible. The coefficient on price (α) is imprecisely estimated, but has the right sign. Market size increases with population and marginal costs are increasing in distance. The signs of the historical cost variables are correct, but insignificantly different from zero.

The coefficients on airport presence are significantly different from zero on both the demand and cost side. The mean valuation of product quality increases with the airport presence variables, with presence at the origin city having a larger effect than presence at the destination city. In addition, serving a large number of routes appears to lower costs. I am not sure how to explain the asymmetry in costs between origin and destination cities; it may be the result of model misspecification, so that the demand structure does not capture all of the asymmetry.

The variance of the market-specific demand error ξ is large, indicating that it would be inappropriate to treat price exogenously. However, the imprecise estimation of α indicates a need for better instruments on the cost side.

IV. Conclusion

The preliminary structural estimates in Table 2 are from an overly simplified model of product differentiation. However, they do account for many of the difficulties found in previous work; they represent some improvement over simple logit models of airline demand. Future work will have to develop even richer demand structures (that could, for example, incorporate price discrimination) and will have to account for endogenous entry along with equilibrium prices and quantities. The single-period profit model here would ideally be embedded in a dynamic context. The profit model of this paper is, for example, consistent with the general dynamic framework of R. Ericson and Pakes (1989).

The results in Section III are challenging in that they imply that airport presence is related both to cost and to demand advantages. This has several implications for policy discussions in the regulated airline industry. The coincidence of price increases with cost decreases makes welfare analysis particularly difficult. This difficulty is compounded by the socially ambiguous nature of much of the product differentiation in this industry; the area under an airline's differentiated product demand curve should not be confused with social benefit. While an increase in demand following from more frequent flights may represent an increase in social welfare, it is much more difficult to associate frequent flyer plans with such increases.

REFERENCES

Berry, S., "Estimation of a Model of Entry in the Airline Industry," Yale University Working Paper, 1989.

Borenstein, S., "Hubs and High Fares: Dominance and Market Power in the U.S. Airline Industry," *Rand Journal of Economics*, Autumn 1989, *20*, 344–65.

______, "The Dominant-Firm in Multi-Product Industries: Evidence from the U.S. Airlines," *Quartely Journal of Economics*, forthcoming 1990.

Caves, D. W., C., Laurits R. and Tretheway, M. W., "Economies of Density versus Economies of Scale," *Rand Journal of Economics*, Winter 1984, *15*, 471–89.

Ericson, R. and Pakes, A., "An Alternative Theory of Firm and Industry Dynamics," Columbia University working paper, 1989.

Levine, M., "Airline Competition in Deregulated Markets: Theory, Firm Strategy and Public Policy," *Yale Journal on Regulation*, Spring 1987, *4*, 393–494.

McFadden, D., "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, September 1989, *57*, 995–1026.

Morrison, S. and Winston, C., "Enhancing the Performance of the Deregulated Air Transportation System," *Brookings Papers on Microeconomics*, 1:1989, 61–112.

Pakes, A. and Pollard, D., "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, September 1989, *57*, 1027–58.

Reiss, P. and Spiller, P., "Competition and Entry in Small Airline Markets," University of Illinois Discussion Paper, 1987.

Sattinger, M., "Value of an Additional Firm in Monopolistic Competition," *Review of Economic Studies*, April 1984, *51*, 321–32.

Airline Mergers, Airport Dominance, and Market Power

By SEVERIN BORENSTEIN*

Although the laissez-faire antitrust policies of the 1980s affected market structure in many U.S. industries, the effects have probably been most visible in the airline industry. In the last few years, there has been increasing focus on the market power that an airline may obtain when it offers most of the service from an airport. This power may result not just from the ability to exclude competitors from city-pair markets, but also from the advantage that a dominant airline could have when competitors are present. For instance, the dominant airline in an area may be able to utilize marketing devices such as frequent flyer plans and travel agent commission overrides (programs that reward agents for directing a high proportion of their business to the airline) more effectively than other airlines in order to attract travelers who originate their trips in that area.[1]

This paper analyzes the effects of two controversial airline mergers that resulted in airport dominance and may have created substantial market power. Both acquisitions occurred in the fall of 1986—Northwest's merger with Republic Airlines and Trans World Airlines' (TWA) purchase of Ozark Airlines. In each case, the merger left the surviving carrier with more than three-quarters of the traffic at the major hub airport where the merging airlines had competed, Minneapolis/St. Paul (MSP) for Northwest and St. Louis (STL) for TWA.

I. History of the Mergers

Prior to their merger, Northwest (NW) and Republic (RC) airlines served many of the same routes from MSP and were the most likely potential competitors for one another when only one of them served a MSP route. They were by far the two largest carriers at MSP, with NW accounting for 42 percent of the enplanements there and RC getting 37 percent. The same basic description applies to the merger of TWA with Ozark (OZ) and its effect in St. Louis, though the carriers' positions were not quite as symmetric. TWA had carried 57 percent of the traffic from STL and OZ 25 percent. Both MSP and STL received some service from other carriers, but it was limited in most cases to flights from the other airlines' hub airports. United, for instance, flew nonstop to MSP both before and after the merger, but only from Denver and Chicago, its two primary hubs.

Preliminary discussions of each of the mergers began in 1985 and final agreements were completed during the first quarter of 1986, at which point permission for each merger was requested from the U.S. Department of Transportation (DOT). The NW/RC application was granted in August of 1986 and the TWA/OZ application in September. Both mergers took place in October 1986.

II. The Evidence on Price Changes

One obvious test of the acquisition of market power is a before and after comparison of the merging firms' prices. Table 1 presents the prices of the eight largest domestic carriers on routes that include their major hubs *relative* to the industry average prices for routes of the same distance.[2] The data are from the third quarters of 1985, 1986, and 1987. The first of these periods is probably before or during the preliminary merger negotiations. The second is after

---

*Department of Economics, University of California, Davis, CA 95616. I thank Sarah Lane and Greg Werden for helpful comments, Janet Netz for excellent research assistance, and the National Science Foundation for research support.

[1] See Michael Levine (1987) and my 1989 paper for a description of airport dominance advantages.

[2] For each route, the ratio of the carrier's average price on the route to the industry average price for similar distance routes is calculated. Each route is weighted by the passenger-miles that the carrier provided on the route.

TABLE 1—AVERAGE PRICES OF MAJOR AIRLINES AT THEIR PRIMARY HUBS

Airline	Airport	Relative Prices[a] 1985	1986	1987	Change[a] 85–87
Texas Air	Denver	−19.4	−20.6	−4.3	18.7
NW/RC	Minneapolis	12.4	10.0	25.2	11.4
TWA/OZ	St. Louis	11.0	7.1	20.5	8.6
Piedmont	Charlotte	14.3	26.0	24.0	8.5
American	Dallas (DFW)	20.0	22.5	26.9	5.8
Delta	Atlanta	26.2	37.8	24.4	−1.4
USAIR	Pittsburgh	23.9	17.5	13.1	−8.7
United	Chicago (ORD)	15.8	19.1	1.9	−12.0

[a] Shown in percent.

agreement on the mergers and just before they were consummated. The third is nearly a year after the mergers took place. The table shows that, with the exception of Texas Air's increases of the depressed prices at Denver (following the demise of Peoples Express and Frontier), the two mergers examined here accounted for the largest hub-airport price increases during this time. Minneapolis and St. Louis were two of the least expensive hubs to travel to or from in 1985 (with prices of the dominant airlines at these airports just 11–12 percent above the national average prices), but by 1987 were about as expensive as the other major dominated hubs.

The aggregation done to produce Table 1 weights each route by the number of passenger miles that the carriers provided on it. Though this is appropriate in gauging the impact of price changes on aggregate consumers' welfare, it is not appropriate for a cross-sectional comparison of market power. Table 2 presents unweighted average price changes on routes from the hubs (relative to the change in average industry prices for similar distance routes) and divides the markets into four categories based on the market structure before and after the merger. The routes are partitioned into those on which only one of the merging carriers was an active competitor prior to the merger and those on which both merging carriers competed. Each of these groups is then divided into routes with at least one other competitor and those without another airline.[3]

TABLE 2—MERGING AIRLINES' PRICES AT THEIR PRIMARY HUBS

	Other Firms	Mkts	Relative Prices[a] 1985	1986	1987	Av. Change[a] 1985–87
NW & RC	Yes	16	3.1	0.2	10.1[d]	6.7
			(2.8)	(4.5)	(5.9)	(4.3)
NW or RC	Yes	41	14.3[b]	21.2[b]	19.9[b]	6.0[c]
			(2.6)	(3.5)	(2.8)	(2.6)
NW & RC	No	11	15.2[d]	32.1[b]	37.8[b]	22.5[b]
			(8.2)	(10.3)	(7.5)	(5.2)
NW or RC	No	16	27.0[b]	36.6[b]	39.4[b]	12.0[c]
			(6.7)	(9.5)	(7.1)	(5.5)
Total		84	14.7[b]	21.5[b]	24.1[b]	9.5[b]
			(2.3)	(3.3)	(2.7)	(2.1)
TWA & OZ	Yes	19	−1.3	−2.7	3.2	4.6
			(6.1)	(4.0)	(4.6)	(7.5)
TWA or OZ	Yes	29	10.5[c]	4.7	5.7	−3.0
			(4.0)	(4.2)	(4.4)	(3.1)
TWA & OZ	No	9	39.6[b]	55.5[b]	27.4[b]	−5.8
			(7.5)	(13.2)	(2.4)	(6.4)
TWA or OZ	No	10	56.0[b]	61.4[b]	33.5[b]	−12.3[c]
			(12.0)	(11.8)	(8.1)	(4.0)
Total		67	17.8[b]	17.9[b]	12.1[b]	−0.0
			(4.0)	(4.6)	(3.0)	(3.5)

[a] Shown in percent.
[b] Significant at 1 percent level (two-tailed test).
[c] Significant at 5 percent level (two-tailed test).
[d] Significant at 10 percent level (two-tailed test).

The first conclusion from Table 2 is that it provides no evidence that the TWA/OZ merger had a systematic impact on these carriers' prices on St. Louis routes. In fact, their price changes on the 67 STL routes observed averaged almost exactly the same as the industry average price changes. The discrepancy between this result and the impact calculated in Table 1 is accounted for by the large price increases on some of the highest density STL routes. The apparent absence of a strong effect from the merger, however, does not imply that TWA/OZ lacked substantial market power. While the merging carriers' prices stayed near industry average on routes with other carriers, prices were consistently and substantially above industry average where no other carriers were present. Still, even these prices fell relative to the industry average around the time of the merger, significantly so for the routes that had been served by only one of the merging airlines.

---

[3] Active competitor is defined to be a market share of at least 10 percent. Routes with at least 10 passengers per day are included. Price change is measured as the ratio of relative price in 1987:3 to relative price in 1985:3, so the average change is not equal to the change in the average relative price. All route data count only local traffic, not through or connecting passengers.

In contrast, analysis of NW/RC price changes indicates that this merger increased market power at Minneapolis. In all four categories, the merging carriers' prices increased faster than industry average, with the largest average (relative) increase coming on the routes in which the merging carriers had been the only two competitors. Prices on these routes had been substantially above the industry average prior to the merger and they increased to average 38 percent above the industry level by 1987.

The statistically significant 12 percent average (relative) increase in prices on routes that were NW or RC monopolies before the merger as well as after gives support to one or both of two hypotheses about the advantages of a dominant airline at an airport: 1) that airport dominance allows use of marketing devices that increase the attractiveness of a carrier, and 2) that airport dominance lessens the threat from potential competition. The timing of the price increase in this category (coming predominantly before the merger, during the time of negotiations and agreement) lends greater support to the potential competition theory.

The increase in NW/RC prices on routes that had another competitor is less impressive, but the timing of these changes is quite interesting. In fact, only in the first category of markets shown in the table, where both of the merging carriers competed along with at least one other airline, was there a significant average price increase between the third quarters of 1986 and 1987—averaging 10.5 percent (standard error = 4.0 percent). In the other three categories, significant price increases occurred only between the third quarters of 1985 and 1986. The result that routes served by only NW and/or RC experienced price increases primarily before the merger, but price increase on routes with both merging carriers and at least one other airline occurred only after the merger gives support to the conclusion of David Graham et al. (1983) that airlines find it much more difficult to tacitly collude in markets with three carriers than in markets with two carriers.[4]

Not shown in Table 2 are the price changes of competing airlines in the first two categories of routes. For both merger cases examined here, the other airlines on these routes displayed average price increases very close to those of the merging airlines, indicating that the merger did not substantially change the premium that the dominant and merging carriers could charge over its competitors. Still, for both NW/RC at MSP and TWA/OZ at STL, a significant price premium existed, with the dominant airlines' prices averaging about 10 percent greater than those of other airlines.

III. The Evidence on Market Shares

Another approach to diagnosing market power that results from airport dominance is suggested in my 1990 paper, and focuses on market shares. If TWA's dominance of the St. Louis airport increases its attractiveness to people who live in St. Louis, then on the St. Louis-Pittsburgh route, for example, one would expect TWA to capture a higher proportion of the traffic that originates in St. Louis than originates in Pittsburgh. In fact, on this route, TWA/OZ got 57 percent of the St. Louis-originating traffic and 52 percent of the Pittsburgh-originating traffic during the third quarter of 1985. By third quarter 1987, this difference had grown to 61 percent vs. 51 percent. Thus, the difference in TWA's route share by point of passenger origin had increased by 5 percentage points. These differential route shares by point of origin are unlikely to be explained by quality or price differences, because passengers who fly a given route with a given carrier generally face the same prices and service quality regardless of which end of the route is their origination point.

Table 3 presents the average change in the merging carrier's difference in route shares by point of origin for routes on which competition existed both before and after the

---

[4] The difference in changes between the first and third categories in Table 2 is statistically significant for the 1985–86 period, but not for the 1986–87 period.

TABLE 3—AVERAGE DIFFERENCE IN ROUTE SHARE CHANGES BY POINT OF ORIGIN

	All Routes[a]			Direct Routes[a]		
	Mkts	1985–86	1986–87	Mkts	1985–86	1986–87
TWA/OZ at STL	48	0.8	5.0[b]	16	−3.4	6.6[c]
		(1.3)	(1.5)		(2.5)	(2.9)
NW/RC at MSP	57	0.8	1.7	7	0.1	4.5[d]
		(1.7)	(1.7)		(1.9)	(2.1)

[a,b,c,d] See Table 2.

mergers. In addition, the table presents similar results for direct passengers for the subset of these routes on which at least 80 percent of the passengers traveled direct. The results speak fairly clearly: there is no apparent increase in attraction of the dominant airlines before the mergers (third-quarter 1985 to third-quarter 1986), but there does seem to be an increase in the difference in route shares by point of origin in the year following the mergers. On St. Louis routes that TWA/OZ served with other competitors, the merging airlines' net change in share of passengers was 5.0 percent greater for travelers who originated at St. Louis than for those who originated at the other endpoint of these routes. On the direct routes, the difference was 6.6 percent. The figure is roughly comparable for NW/RC on its direct routes from Minneapolis, but is much smaller and statistically insignificant for the sample of all routes it served competitively from MSP. The sign of these route-share changes is consistent with increased attractiveness of the merged carrier to passengers who originate at its dominated airport. The timing of these changes supports the hypothesis that they may be due to frequent flyer programs and other marketing devices. The ability to use such marketing programs was probably enhanced by the mergers only after the mergers actually took place and the programs of the merging companies were combined.

IV. The Evidence on Changes in Service

Studies both before and after airline deregulation have found a strong positive correlation between concentration on a route and load factors (the proportion of seats filled). It appears that prior to deregulation when airlines could not adjust prices, they instead competed by adding flights on a route, which tended to drive down load factors. This sort of competition has decreased, but has persisted since deregulation.

TABLE 4—MERGING CARRIERS' CHANGES IN LOAD FACTORS AND AVAILABLE CAPACITY AT THEIR PRIMARY HUBS

	Other Firms	Mkts	Capacity Change[a] 1986–87	Relative Load Factors[a] 1986	Relative Load Factors[a] 1987	Rel. L.F. Change[a] 1986–87
NW & RC	Yes	6	−28.3[b]	−19.4[b]	0.8	12.8[b]
			(5.7)	(5.0)	(1.9)	(3.2)
NW or RC	Yes	3	−6.2	1.6	−4.1	−3.4
			(22.0)	(6.2)	(7.1)	(4.5)
NW & RC	No	19	−14.4[b]	−10.0[b]	5.9[b]	10.1[b]
			(5.0)	(2.9)	(2.0)	(1.4)
NW or RC	No	27	53.0[b]	−5.7[c]	0.6	3.3[d]
			(15.4)	(2.7)	(2.2)	(1.9)
TWA & OZ	Yes	9	−23.9[b]	−0.4	0.3	0.4
			(4.0)	(5.5)	(2.7)	(3.4)
TWA or OZ	Yes	6	3.6	−17.7[d]	1.2	11.0[d]
			(6.8)	(9.1)	(5.6)	(5.4)
TWA & OZ	No	19	−25.5[b]	−5.6[d]	3.8[d]	5.4[b]
			(4.1)	(2.9)	(2.0)	(1.8)
TWA or OZ	No	39	18.2[d]	−8.0[b]	−1.8	3.6[c]
			(9.9)	(2.6)	(1.8)	(1.5)

[a,b,c,d] See Table 2.

Competition in capacity implies that the mergers examined here would result in a *decrease* in the capacity that the merged carrier would offer on routes where the merging airlines had competed with one another, and an *increase* in average load factors on such routes. The implications for routes on which only one of the merging carriers had operated is less clear. Table 4 examines nonstop routes from the dominated airport on which at least one of the merger partners had offered capacity of at least 110 seats, about 1 flight per day, prior to the merger. The relative load factors and their changes are calculated comparably to the relative price calculations in Table 2.[5]

---

[5] These data are for the second quarters of 1986 and 1987. Data for the third quarter of 1986 are likely to reflect disruption in the carriers' operations just prior to their mergers. Capacity is measured in available seats. Similar results were found using flights.

Table 4 verifies that there were substantial service cutbacks on routes where both of the merger partners had operated, regardless of whether or not other airlines competed on these routes. For NW/RC these decreases in service were accompanied by substantial increases in load factors, averaging about 11 percent of the (relative) load factors prior to the mergers, or about 6 percentage points. Passengers on these routes, however, did not suffer particularly poor service after the merger, given the amount of traffic. Rather, the load factor increases seem to have been a normalization from the depressed levels that had existed on these routes prior to the merger. On the St. Louis routes where both TWA and OZ had competed, the load factor increases that accompanied service cutbacks were less substantial and less statistically significant.

Defenders of these mergers have pointed out the merging carriers' new nonstop service on some routes from their hubs. Overall, however, capacity at these airports seems to have declined following the merger. NW/RC capacity on flights to and from MSP fell 7.6 percent following the merger, though a 3.3 percent increase in MSP service by other carriers caused the overall change to be a 5.5 percent decrease. At St. Louis, overall capacity declined 11.4 percent with TWA/OZ service falling 11.1 percent and the capacity of all other carriers falling 12.4 percent.

These capacity changes may give a hint at the cause of the small price changes at STL following the merger there. At STL, TWA's capacity decrease was met with decreases by other carriers and overall decreases in passenger traffic. At the same time, load factors at STL increased much less than the national average. Thus, it appears that the weak indications of increased market power following the TWA/OZ merger may be a result of weakening demand for travel to and from St. Louis.

V. Conclusion

The two mergers analyzed here offer a unique opportunity to learn more about the market power and competitive advantage that can result from airport dominance. The effects that I have found from the NW/RC and TWA/OZ mergers are broadly consistent with the conclusions that Gregory Werden et al. (1989) reach using substantially different methods: The evidence of increased market power at Minneapolis following the NW/RC merger is consistent and convincing, but the evidence in the TWA/OZ case is weaker. NW/RC's price increases were significant not just on routes that both airlines had served prior to the merger, but also on routes where only one of the two merger partners competed with another airline or operated without active competition. Furthermore, the timing of many of these price increases implies that a reduction of competition occurred even before the merger was consummated. In contrast, the changes in market share and capacity that might indicate market power did not appear until after the mergers. Theories that explain these differences in the timing of price changes vs. capacity and market share changes remain for future work. Likewise, the distinctly different results from two fairly similar mergers are also worthy of further investigation.

REFERENCES

Borenstein, Severin, "The Dominant-Firm Advantage in Multi-Product Industries: Evidence from the U.S. Airlines,'" *Quarterly Journal of Economics*, forthcoming 1990.

______, "Hubs and High Fares: Dominance and Market Power in the U.S. Airline Industry," *Rand Journal of Economics*, Autumn 1989, *20*, 344–65.

Graham, David R., Kaplan, Daniel P. and Sibley, David S., "Efficiency and Competition in the Airline Industry," *Bell Journal of Economics*, Spring 1983, *14*, 118–138.

Levine, Michael E., "Airline Competition in Deregulated Markets: Theory, Firm Strategy, and Public Policy," *Yale Journal on Regulation*, Spring 1987, *4*, 393–494.

Werden, Gregory J., Joskow, Andrew S. and Johnson, Richard L., "The Effects of Mergers on Economic Performance: Two Case Studies from the Airline Industry," U.S. Department of Justice Discussion Paper No. EAG 89-15, October 1989.

STRIKES, BARGAINING, AND ARBITRATION: NEW DEVELOPMENTS[†]

Can Strategic Bargaining Models Explain Collective Bargaining Data?

By JOHN KENNAN AND ROBERT WILSON*

Much effort has recently been devoted to the development of strategic models of bargaining with private information. At the same time, empirical analyses of data on collective bargaining in the labor market have generated an interesting list of facts to be explained. In this paper we assess the prospective interactions between these theoretical and empirical developments. The style here is expository: detailed arguments have been ruthlessly suppressed. A more complete treatment can be found in our 1989 article.

The theoretical models deal with conflicts between rational bargainers in situations where at least one of the bargainers has private information, as in the case where an employer knows more about the value of labor's product than the workers do. We start by explaining why private information is important in building models of strikes.

In a naive model of collective bargaining, the union starts with a tough demand, and the firm starts with a low offer. As the negotiations continue the union reduces its demand, moving along a downward-sloping "resistance curve," and the firm increases its offer, moving along an upward-sloping "concession curve," until eventually the resistance curve and the concession curve intersect, and a deal is made. John Hicks (1963) pointed out that this kind of model is seriously flawed: once the union and the firm know the model, they can obviously do better; for example, by settling immediately at the wage predicted by the model. Consequently, if a strike occurs, it must be a mistake, or else one must suppose that unions and employers are not rational bargainers. The trouble then is that economics does not have much to say about when strikes are more or less likely to occur: the optimal number of mistakes is zero, and this does not vary as the explanatory variables change.

I. Private Information

Suppose that management is pleading poverty but the union does not believe them. It may be that the union does not know exactly how much profit is available, so they make a tough demand that is refused and they go out on strike. It turns out that the contract is not as profitable as the union had guessed, and they become convinced of this by seeing that the company is willing to wait out a long strike rather than settle at a high wage. Thus, the strike is finally settled at a relatively low wage.

Given that profits were actually low, the union would have been better off accepting the low wage in the first place, without a strike, and this would have left the firm better off as well. Thus, after the fact, the strike can be regarded as a mistake. But, at the beginning, the union knows that this is one of two possible outcomes, and the other outcome is that profits are actually high, and the firm agrees to the tough demand immediately because the opportunity cost of a strike

---

[†] *Discussants*: Peter Cramton, Yale Unversity; Brian McCall, University of Minnesota; David Bloom, Columbia University.

*Department of Economics, University of Iowa, Iowa City, IA 52242, and Business School, Stanford University, Stanford, CA 94305, respectively. Research support from the National Science Foundation and the Hewlett Foundation is gratefully acknowledged.

is too high. The upshot is that a strike can occur even if both sides behave rationally.

The basic reason for conflict can be examined in a simple one-shot model of pie-splitting with private information on one side. Suppose a union and a firm bargain over the division of a pie, defined as the difference between the firm's demand price for labor, and the workers' supply price. The firm knows the size of the pie, while the union knows that the pie is either a big number, B, if the firm is doing well, or a small number, S, if the firm is doing badly; the probability of a big pie is p. If the (risk-neutral) union somehow has the right to make credible all-or-nothing demands, then it will demand B if $pB > S$, and a strike will result if the actual pie is small. It may seem that this outcome arises only when the uninformed bargainer is allowed to dictate the bargaining mechanism. It can easily be shown, however, that the union's choice is an efficient member of the class of all feasible mechanisms. On the other hand, there are other efficient members of this class that do not involve strikes. Efficiency in this context must be defined relative to the information available when the bargainers choose a mechanism: strikes cannot be efficient *ex post*.

From the point of view of economic policy, an important aspect of this problem is the possibility that mediation or arbitration could lead to substantial gains in the efficiency of the bargaining process. At first sight, this may seem trivially true. Since efficient no-strike mechanisms exist, it seems that a mediator could steer the bargainers toward such a mechanism, or that an arbitrator could impose it, without hurting anyone. Surprisingly, this is false, as can be shown by a simple example. Suppose that the big pie is 12 million dollars, the small pie is \$8m, and p is 90 percent. The bargainers might reach an impasse in which the union will either sign a contract which is worth \$8.5m, or else strike long enough to destroy one-fourth of the pie, and then settle for \$5.5m. If the firm is doing well, it will sign the \$8.5m contract and obtain \$3.5m profit, because it can do no better by taking the strike. If the firm is doing badly, it will take the strike, leaving a profit of \$0.5m.

If the firm's private information is unverifiable in this example, there is nothing that a mediator could do that would avoid a strike. The union's expected payoff is \$8.2m, and if the pie is small, the firm will not pay more than \$8m. The mediator might be told by the firm that the pie is small, and could then suggest to the union that it should accept (say) \$7m now, rather than \$5.5m. If the union were foolish enough to accept this, then a firm that is doing well would (confidentially, and without fear of detection) assure the mediator that it is doing badly, and so make off with a profit of \$5m, rather than \$3.5m, and the union's payoff would always be \$7m. Thus the strike can be seen as the result of divergent expectations, since the union expects \$8.2m, the total pie is only \$8m and the firm expects to break even.

II. A Synopsis of the Facts to be Explained

The empirical literature on collective bargaining has established a number of regularities that a successful theory must accommodate. Here are the main points.

1) Strikes are rare. The probability of a strike is about 15 percent, when negotiating a contract to last 3 years. 2) The median strike duration is about 21 days. Although there is an 85 percent chance of settlement before a strike begins, during a strike the probability of settlement is only about 3 percent per day. 3) The settlement probability decreases as the strike goes on. That is, the hazard function is generally downward sloping. 4) The probability of a strike rises when the economy improves, but strike duration falls. Thus strike incidence is procyclical, but duration is countercyclical. 5) There is little apparent relationship between strike duration and the terms of settlement.

III. Attrition Models

The first model considered is the war of attrition, in which compromise is not allowed, and the winner takes the whole pie. The symmetric version of the model analyzed by Paul Milgrom and Robert Weber (1985) assumes that the pie is of known size, π. Each side's cost of delay accumulates at a fixed rate c per time period, and the rates c_u and c_f for the union and the firm are drawn independently and privately from a distribution $F(c)$. If the duration of the strike is $B = \pi/c$, then the winner will just break even, so the union's break-even point is $B_u = \pi/c_u$, and the firm's is $B_f = \pi/c_f$. The optimal strategies are fully characterized by the property that player i continues the struggle as long as the marginal cost c_i is less than the marginal expected benefit πq_j, where q_j is the other player's quit rate (the conditional density of stopping, given that neither player has stopped previously). The lower is the delay cost, the longer the firm will continue, and similarly for the union, so each player's quit rate is decreasing. The hazard rate is the sum of the parties' quit rates, and therefore it also declines with duration, which agrees with the data. Thus

$$c_f(t) = \pi q_u(t), \qquad c_u(t) = \pi q_f(t)$$

$$\pi h(t) = \pi\left[q_u(t) + q_f(t)\right] = c_f(t) + c_u(t)$$

Note that the delay costs do not vary with time, but, in equilibrium, there is a type of the firm that concedes at t, and $c_f(t)$ refers to this firm's delay cost. The model gives a very specific interpretation of the hazard function: the quit rate for each player at time t is proportional to the delay cost of the opponent who would concede at time t. The empirical hazard function can be used to infer the distribution of delay costs.

A basic result of the symmetric attrition model is that, on average, *the duration of the strike is half the distance to the weaker party's break-even point*. In particular, if the two parties happen to have equal costs, then the conflict will consume the entire pie. To take another example, if the distribution of costs is uniform, the joint cost of the strike is half of the pie, on average. This makes it seem as if each negotiation will lead to a long and costly strike, contrary to the data. But, the conflict may be confined to a small part of the available rent, and then negotiations that conclude without a strike can be interpreted as very short strikes where the pie is very small, relative to the costs of delay. In fact, in the U.S. data, strikes are recorded only if they last more than one day, and in any case delay costs surely exist even before the expiration of the old agreement.

IV. Signalling and Screening Models

We consider next a situation in which a monopoly union supplies a fixed amount of labor, with supply price w. The employer's demand price, v, for this labor is private information, and, from the union's point of view, v is a random variable that lies in an interval strictly above w.

It is both a strength and a weakness of strategic bargaining models that very specific assumptions are made about the rules of negotiation: who is allowed to make offers, and when. Two sets of rules have been emphasized in the theoretical literature. In the signalling model, the union makes an initial offer that is accepted if the employer's demand price is high. If the union's offer is rejected, the employer waits for some time so as to signal to the union that the demand price is low, and then makes a single offer that the union accepts. In this model, the employer's offer reveals the demand price, and the settlement splits the difference between this and the union's supply price.

In the screening model, the rules specify that the union makes a sequence of offers, with a fixed time interval between offers, and the firm decides whether to accept the current offer, or wait for a later one. This leads to an equilibrium in which the union can charge a high wage if the demand price is high, while a firm with a low demand price is more patient, and settles at a lower wage after taking a strike.

Table 1 illustrates the signalling and screening models by means of an example where the supply price w is \$10 per hour, while the demand price v has a triangular distribution on the integers from 11 to 20, with weights proportional to 1,2,3,4,5, 5,4,3,2,1. For simplicity, we ignore the union's initial offer in the signalling model, and concentrate on the speed at which the equilibrium sorts out the various types of the firm. The results show that strikes last much too long in the signalling model, even though the assumed discount factor corresponds to

TABLE 1—AN EXAMPLE OF SIGNALLING AND SCREENING EQUILIBRIA

Demand Price	Signalling Wage	Signalling S.D. (weeks)	Screening Wage	Screening S.D. (weeks)
20	15.00	0	11.14	0
19	14.50	19.0	11.14	0
18	14.00	40.0	11.14	0
17	13.50	63.5	11.12	0.8
16	13.00	90.2	11.08	2.0
15	12.50	120.9	11.05	3.3
14	12.00	157.3	11.03	4.5
13	11.50	201.8	11.02	5.7
12	11.00	259.2	11.01	7
11	10.50	340.1	11.00	8

Notes: S. D. denotes Strike Duration.

Firm's Demand Price (dollars per hour)

v	11	12	13	14	15	16	17	18	19	20
Frequency	1	2	3	4	5	5	4	3	2	1

Union's supply price: $w = \$10$ per hour.

Discount factor: .995 per week.

an interest rate of about 2 percent per month, so that both sides are very impatient. The long delay is needed to discourage firms with high valuations from imitating those with low valuations. For example, if v is 20, the firm could pretend that v is 19, waiting 19 weeks to settle at a wage of \$14.50. The present value of profit resulting from this deception would be $.995^{19}$ (\$5.50), which equals the \$5 profit that this firm gets by playing straight. Any delay less than 19 weeks is inconsistent with equilibrium, since a shorter delay makes deception more profitable.

Peter Cramton and Joseph Tracy (1989) have recently argued that the signalling model can be modified so as to fit the data: by assuming a finite contract duration; by increasing the gap between w and v; and by recognizing that in practice work often continues under an extension of the old contract, while negotiations proceed. This depends heavily on the costs attributed to contract extensions. Taking the firm with demand price of \$16 as representative, suppose that the contract lasts 3 years, and further suppose w is 0 instead of \$10 (although this implies an unrealistically high union wage differential). Then, strike duration falls from 90.2 weeks to 22.6 weeks, which is still much too long to fit the data.

Table 1 shows that the strike durations implied by the screening model are, if anything, too short relative to the data discussed above, but strikes may last as long as 7 or 8 weeks, so that with some modification of the parameters reasonable explanations of duration might be obtained. In this model, the union is forced to accept a wage that is close to the lower bound of v. Of course, the union captures everything between w and the lower bound of v, but little more. This seems consistent with observed union wage differentials.

The screening model must admit unobserved heterogeneity in order to escape the implication that the union's offer is accepted with probability .85, while the acceptance probability for the second and later offers drops to about .03 (assuming that the commitment period is one day). The simplest way to introduce heterogeneity is to assume that each observation is drawn from one of two populations, where each population represents a model with fixed parameters, and a specified distribution of private information. An immediate implication is that the (average) hazard function will tend to decline, because the population with the lower settlement rate will come to dominate the mixture of continuing strikes as time goes on.

V. The Coase Property

An intriguing feature of the screening model is that it can potentially explain the observation that strike incidence and strike duration move in opposite directions over the business cycle. The explanation, following Faruk Gul et al. (1986), rests on an admittedly arbitrary assumption that the union can sustain each offer for a longer period of time during a recession, before receding to its next offer. This has the direct effect of stretching out the negotiations for a longer period of time, which explains why strikes are longer during recessions. But, there is also an indirect effect, in that the firm is more willing to accept each offer when it knows that the union's next offer will be slow in coming. The indirect effect is of second-order in explaining strike duration, but it explains why the union's first offer is more likely to be accepted in a recession,

TABLE 2—THE COASE PROPERTY[a]

	Commitment			
	1 Week		1 Day	
Demand Price	Wage	S.D. (weeks)	Wage	S.D. (days)
20	11.14	0	11.03	0
19	11.14	0	11.03	0
18	11.14	0	11.03	0
17	11.12	0.8	11.03	1.0
16	11.08	2.0	11.02	1.6
15	11.05	3.3	11.01	3.4
14	11.03	4.5	11.01	4.5
13	11.02	5.7	11.00	5.7
12	11.01	7	11.00	7
11	11.00	8	11.00	8
Mean	11.07	2.9	11.02	2.9
Strike Incidence:[b]	.77		.80	

Notes: See Table 1.

[a] Discount factor: .999 per day.

[b] Shown in percent.

and thus it explains the observation that strike incidence is procyclical.

In the example of Table 2, suppose that offers are made once a day, instead of once a week, and replace the weekly discount factor of .995 with a daily discount factor of .999, which is roughly comparable. The table shows that one effect of speeding up negotiations is that wage settlements are less favorable to the union, but since the wage offers were so low to begin with, this does not seem important. The shorter commitment period has little effect on the probability that offers are rejected, so that strikes last about the same number of periods, and since a period is now a day rather than a week, strikes are much shorter in real time. Strike incidence rises, but only a little: from 77 to 80 percent. Thus, although the Coase property can explain why the incidence and duration of strikes might move in opposite directions, it remains to be seen whether it is capable of explaining the magnitudes of these movements in the data.

REFERENCES

Cramton, Peter C. and Tracy, Joseph S., "Strikes and Delays in Wage Bargaining: Theory and Data," unpublished manuscript, Yale University, December 1989.

Gul, Faruk, Sonnenschein, Hugo and Wilson, Robert, "Foundations of Dynamic Monopoly and the Coase Conjecture," *Journal of Economic Theory*, June 1986, *39*, 155–90.

Hicks, John R., *The Theory of Wages*, 2nd ed., New York: Macmillan, 1963.

Kennan, John and Wilson, Robert, "Strategic Bargaining Models and Interpretation of Strike Data," *Journal of Applied Econometrics*, December 1989, Suppl., *4*, S87–S130.

Milgrom, Paul R. and Weber, Robert J., "Distributional Strategies for Games with Incomplete Information," *Mathematics of Operations Research*, November 1985, *10*, 619–32.

Strikes and Bargaining: A Survey of the Recent Empirical Literature

By David Card*

Although strikes have captured the attention of economists for many years, the microeconometric analysis of collective bargaining disputes is still in its infancy. Until recently, a major stumbling block was the absence of data on contract negotiations and strikes. Within the past decade, however, researchers have assembled a series of large-scale micro data sets that combine information on contract expirations, wage settlements, and work stoppages. There is now a significant body of work quantifying the relations between strike frequency and duration, wage settlements, product and labor market conditions, and other aspects of the bargaining environment.

This paper presents a selective survey of the new microeconometric literature on collective bargaining. Rather than provide a comprehensive catalogue of findings, I highlight some of the themes that have attracted the attention of empirical analysts, including the cyclical responsiveness of strike activity and the relation of wages to strike outcomes. As I hope to show, the data collection effort that underlies this recent work has paid off handsomely. In less than a decade, the breadth and depth of knowledge on collective bargaining and strikes has improved dramatically.

I. The Theoretical Background

Microeconometric studies of strikes and bargaining have been influenced by three sets of ideas: the model of Orley Ashenfelter and George Johnson (1969); the "joint-cost" hypothesis articulated by John Kennan (1980) and Melvyn Reder and George Neumann (1980); and the asymmetric information models developed by Beth Hayes (1984), Sanford Morton (1983), Joel Sobel and Ichiro Takahashi (1983), and Drew Fudenberg and Jean Tirole (1983). The importance of Ashenfelter and Johnson's paper is twofold. On the theoretical side, it focuses explicitly on the lack of information among union members. In this regard, it lays the groundwork for later one-sided asymmetric information models. On the empirical side, Ashenfelter-Johnson were the first to attempt a systematic regression analysis of strike data, and they provided a particularly durable hypothesis linking strike probabilities to previous real wage changes.

The joint-cost hypothesis states that, whatever the mechanism generating disputes, the probability of a strike and the expected duration of a work stoppage will be lower, the higher the *joint* cost of strike to the firm and its employees. This simple hypothesis provides an interesting perspective on the effects of the business cycle on strikes. Whereas improving cyclical conditions in the product market increase the joint costs of a strike, and thereby reduce strike propensities, improving cyclical conditions in the external labor market reduce the opportunity costs of a stoppage, and thereby increase dispute rates. Following this insight, many recent studies of strike activity have attempted to control for both firm-level product demand and local labor market conditions.

One-sided asymmetric information models combine elements of both the Ashenfelter-Johnson model and the joint-cost hypothesis. According to the usual formulation of these models, some component of profitability is unobservable to union members. A strike is viewed as a mechanism that allows workers to extract higher wages from more profitable employers (see Kennan, 1986). Ashenfelter and Johnson's downward-sloping concession schedule is deduced as a direct implication of incentive compatibility.

---

*Princeton University, Princeton, NJ 08544.

In contrast to Ashenfelter-Johnson, however, the profitability information revealed by a strike is not otherwise observable. The implications of the joint-cost hypothesis survive, since increases in the joint costs of a strike reduce the likelihood that a strike mechanism is selected, and similarly reduce the probability and expected duration of a strike, given that a strike mechanism is selected.

II. Strike Incidence

A first-order fact about private sector bargaining in North America is that 10–15 percent of contract negotiations involving relatively large numbers of workers lead to a work stoppage. Two U.S. studies with broad industry coverage (Cynthia Gramm, 1987; Sheena McConnell, 1987) estimate the average probability of strikes among contract negotiations with 1000 or more workers at 13 percent. A Canadian study by Morley Gunderson et al. (1986), covering contract negotiations with 200 or more workers, estimates a similar 13 percent dispute rate. The incidence of strikes is slightly higher in manufacturing industries, and also varies substantially across narrowly defined manufacturing industries.

As one might expect from the earlier literature devoted to explaining the aggregate *number* of strikes in a year, there is considerable year-to-year variation in the average probability of disputes. Rates in both the United States and Canada ranged from 10 percent to as high as 35 percent in the decade of the 1970s. A traditional explanation for the time-series variation in strike activity is the state of the business cycle. Indeed, this is one of the most widely investigated hypotheses in the recent literature.

There is systematic evidence that higher unemployment reduces the likelihood of strikes. Perhaps the strongest evidence comes from the study by Susan Vroman (1989), that follows contract negotiations of 250 bargaining pairs over the period 1957–82. Her estimates indicate a 2–3 percentage point reduction in the probability of strikes for each 1 percentage point increase in the prime-age male unemployment rate. Similar estimates are reported by Gunderson et al. and myself (1990) using Canadian data, while somewhat smaller (but still negative and statistically significant) estimates emerge from the study by McConnell (1987). Although not strictly comparable, the estimates in Joseph Tracy (1986) suggest a slightly larger negative effect.

There is much less agreement across studies on the effect of industry-level demand conditions. Tracy (1986) measures demand by the deviation of industry employment from its trend. His estimates suggest that higher industry employment reduces the probability of strikes. John Abowd and Tracy's (1989) analysis of a very similar specification, however, suggests an effect of the *opposite* sign. McConnell (1987) regresses monthly industry-average strike probabilities on an index of the real industry selling price. Her results indicate that an increase in an industry's relative output price leads to a decrease in the probability of a strike. This contrasts to the findings for Canadian manufacturing industries in my paper (1990), which suggest a *positive* association between the industry value-added price index and the probability of disputes.

The conflicting evidence on the effects of industry-level demand suggests the need for caution in the interpretation of the more systematic effects of unemployment. The latter are often taken as evidence in favor of the joint-cost hypothesis, on the assumption that higher unemployment reduces the earnings opportunities of striking workers. But, the joint-cost hypothesis also predicts a relation between dispute rates and the state of firm or industry-specific product demand. An alternative hypothesis is that external labor market conditions affect the union's ability to withhold labor services. Existing bargaining models abstract from the threat of strike breakers and dissenters within the union. Nevertheless, union control over the supply of labor is limited. (See the recent paper by John Schnell and Gramm, 1989.) If the union's ability to maintain an effective picket line is reduced in a depressed labor market, one would expect a moderating effect of higher unemployment on dispute rates, apart from any effect via the joint cost of strikes.

An alternative explanation for the changing propensity to strike concerns recent wage and price movements. Ashenfelter-Johnson found that nominal wage and price increases in the previous 2 years affect the aggregate number of strikes in a quarter with equal and opposite elasticities. Several recent authors have confirmed the correlation between dispute rates and contract-level changes in real wages. Gunderson et al. find that strike probabilities are negatively associated with the change in real wage rates during the term of the expiring agreement. Vroman finds that strike probabilities are negatively related to the change in contract wages relative to the change in aggregate wages during the contract period. The results in my 1990 paper suggest a similar conclusion: a higher relative wage at the end of the expiring agreement reduces the probability of strikes.

Whether in relative or absolute terms, these estimates suggest that there is a systematic effect of the wage rate inherited from the previous contract. This finding is difficult to reconcile with existing one-shot bargaining models, and underscores the point that contract negotiations typically proceed in the shadow of an existing agreement. Recent work by Peter Crampton and Tracy (1989) extends the asymmetric information model to allow for continuing negotiations without a work stoppage under the terms of the old contract. Their initial results suggest that the level of real wages inherited from the previous contract governs the choice between a work stoppage and continued negotiations. Given the evidence, this is clearly a fruitful line of inquiry.

III. Strike Duration

The duration of work stoppages have been less studied than strike incidence, in part because of data limitations: even large samples of contract negotiations yield relatively few strikes; and the major source of contract data for Canada provides no information on strike length. The analysis of strike durations has been aided, however, by the availability of comprehensive lists of strikes for both the United States and Canada. Although these lists contain very little information on the associated contract negotiations, they provide a useful supplement to the available contract-based data sets.

Estimates of mean and median strike duration suggest that the typical work stoppage is short: mean durations of strikes in larger manufacturing contract negotiations range from 40 to 50 days in the United States and Canada; median durations range from 25 to 40 days. As a fraction of mean contract duration, these translate into losses of approximately 5 percent of working time, conditional on a strike, and 0.5 to 1.0 percent, conditional on a contract negotiation. Expected durations of strikes vary substantially across industries, and are positively associated with the probability of strikes (see my 1988 paper).

As with strike incidence, the time-series variation in strike duration has been traditionally explained by reference to the business cycle. And, as with strike incidence, the evidence on the cyclicality of strike durations is unclear. Following the insight of the joint cost hypothesis, Tracy (1986, 1987), McConnell (1987), and I (1990) attempt to isolate the separate effects of cyclical labor market conditions and cyclical product market conditions. Tracy's estimates show a strong negative effect of the state employment growth rate on the expected duration of disputes, but a negligible effect of the industry growth rate. Using much larger samples of strikes, McConnell and I both find weak negative effects of unemployment rates on durations, but relatively large and statistically significant positive effects of industry selling prices on expected strike durations.

This positive association seems to contradict the evidence in Kennan (1985) and Alan Harrison and Mark Stewart (1989) of a *negative* relation between strike durations and the index of industrial production. Harrison-Stewart report a roughly 14 percent increase in strike durations from peak to trough of the production index, while Kennan reports a 35 percent increase. The discrepancy may arise in part from the choice to measure demand with prices or quantities. It also may be related to the distinction between aggregate and industry-specific demand. Neither Kennan nor Harrison-Stewart attempt to separate the effects of industry-specific demand conditions from those of the aggregate economy. In fact, Harrison-Stewart find a much weaker relation between strike durations and industry-specific output than economywide output.

One aspect of strike durations that has received some attention is the time profile of conditional settlement rates. Kennan (1980) initially proposed to evaluate the joint-cost hypothesis using evidence on the shape of the settlement hazard function over time. More recently, Kennan and Robert Wilson (1990) have proposed to test and compare alternative private information and war-of-attrition models of strikes using this same information. Any such test is complicated by the problem of distinguishing true duration dependence from the effects of unobserved heterogeneity. The evidence to date on the shape of the hazard function is mixed. Kennan (1985) finds that conditional settlement rates first decrease, then increase. By comparison, Harrison-Stewart and Gunderson and Angelo Melino (1988) estimate monotonically decreasing hazards.

IV. Strike and Wages

A key prediction of the Ashenfelter-Johnson model, and later one-sided asymmetric information models, is that negotiated wage outcomes and strike durations are negatively correlated. An early test of this prediction was conducted by Henry Farber (1978) in a highly structured analysis of some 150 contract negotiations. Farber's estimates suggest that the concession schedule is downward sloping, at a rate of about 6 percent per year.

The first attempt to relate wage settlements and strike outcomes using a broad cross-section of wage settlements is W. Craig Riddell's (1980) study, that uses Canadian contract data from 1953 to 1973. Contrary to the predictions of the Ashenfelter-Johnson model, Riddell found that strikes are associated with significantly *higher* wage increases. Robert Lacroix (1986) presents a similar specification that confirms the positive association between wage settlements and strike outcomes, controlling for expected inflation and the current unemployment rate. Lacroix points out, however, that the positive relation between wage increases and strike outcomes does not survive an unrestricted treatment of the year effects in the wage change equation. Apparently, the positive association of wages and strike probabilities is at the aggregate level, rather than at the individual contract level. In this regard it is interesting to note that industry average strike probabilities are also positively correlated with wage rates.

Two recent studies (by McConnell, 1989, and myself, 1990) set out to test the prediction of a negative relation between strikes and wages on U.S. and Canadian contract data. In contrast to earlier studies that model nominal wage changes, both McConnell and I model the expected average real wage rate during the term of a contract. The wage equations in both studies include unrestricted year effects and measures of comparison wages, unemployment rates, and industry demand. Despite these similarities, the two studies lead to different conclusions. Whereas McConnell finds a significantly negative association between strikes and real wages, my study found no significant relation.

One difference between the studies is their industry coverage: McConnell's data include manufacturing and nonmanufacturing contracts, whereas mine was restricted to manufacturing industries. Indeed, McConnell finds somewhat more negative effects of strikes on nonmanufacturing wage rates. The estimates in my study do suggest lower wage rates after very long strikes (150 days or more). For the vast majority of strikes, however, the Canadian data give no indication of a negative relation between strikes and wages. The difference between these two studies is a puzzling feature of the recent literature that needs to be resolved in order to facilitate further progress.

V. Conclusions

Although the microeconometric study of strikes and bargaining is still in its infancy, a number of substantive conclusions have emerged. First, the probability and intensity of labor disputes vary systematically across industries. Second, changes in the likelihood of a strike for a given bargaining pair are related to observable economic conditions: higher unemployment and real wage gains during the previous agreement reduce the probability of strikes. Third, the average duration of disputes varies with the aggregate cycle. There are also a wide range of findings relating strike behavior to contract length, the seasonality of contract expirations, and uncertainty that have not been surveyed here. Nevertheless, many puzzles remain. The effects of firm-specific product demand on the probability and expected duration of strikes are unclear, as is the association between negotiated wage outcomes and strike activity. These issues will surely receive further attention in the next decade of research on the microeconometrics of strikes and bargaining.

REFERENCES

Abowd, John and Tracy, Joseph, "Market Structure, Strike Activity, and Union Wage Settlements," *Industrial Relations*, Spring 1989, *28*, 227–50.

Ashenfelter, Orley and Johnson, George, "Bargaining Theory, Trade Unions, and Industrial Strike Activity," *American Economic Review*, March 1969, *59*, 35–49.

Card, David, "Longitudinal Analysis of Strike Activity," *Journal of Labor Economics*, April 1988, *6*, 147–76.

______, "Strikes and Wages: A Test of an Asymmetric Information Model," *Quarterly Journal of Economics*, forthcoming 1990.

Crampton, Peter and Tracy, Joseph, "Strikes and Delays in Wage Bargaining: Theory and Data," unpublished paper, Department of Economics, Yale University, October 1989.

Farber, Henry, "Bargaining Theory, Wage Outcomes, and the Occurrence of Strikes," *American Economic Review*, June 1978, *68*, 262–71.

Fudenberg, Drew and Tirole, Jean, "Sequential Bargaining with Incomplete Information," *Review of Economic Studies*, April 1983, *50*, 221–47.

Gramm, Cynthia, "New Measures of the Propensity to Strike During Contract Negotiations, 1971–1980," *Industrial and Labor Relations Review*, April 1987, *40*, 406–17.

Gunderson, Morley and Melino, Angelo, "The Effects of Public Policy on Strike Duration," Department of Economics Working Paper No. 8807, University of Toronto, April 1988.

______, **Kervin, John and Reid, Frank,** "Logit Estimates of Strike Incidence from Canadian Contract Data," *Journal of Labor Economics*, April 1986, *4*, 257–76.

Harrison, Alan and Stewart, Mark, "Cyclical Fluctuations in Strike Durations," *American Economic Review*, September 1989, *79*, 827–41.

Hayes, Beth, "Unions and Strikes and Asymmetric Information," *Journal of Labor Economics*, January 1984, *2*, 57–83.

Kennan, John, "Pareto Optimality and the Economics of Strike Duration," *Journal of Labor Research*, Spring 1980, *1*, 77–94.

______, "The Duration of Contract Strikes in U.S. Manufacturing," *Journal of Econometrics*, April 1985, *28*, 5–28.

______, "The Economics of Strikes," in O. Ashenfelter and R. Layard, eds., *The Handbook of Labor Economics*, Vol. 2, New York: North-Holland, 1986.

______ **and Wilson, Robert,** "Strategic Bargaining Models and Interpretation of Strike Data," *Journal of Applied Econometrics*, forthcoming 1990.

Lacroix, Robert, "A Microeconometric Analysis of the Effects of Strikes on Wages," *Relations Industrielles*, Winter 1986, *41*, 111–26.

McConnell, Sheena, "Cyclical Fluctuations in Strike Activity," Princeton University Industrial Relations Section Working Paper No. 215, January 1987.

______, "Strikes, Wages and Private Information," *American Economic Review*, September 1989, *79*, 801–815.

Morton, Sanford, "The Optimality of Strikes in Labor Negotiations," Tulane University Discussion Paper No. 83/7, 1983.

Reder, Melvin and Neumann, George, "Conflict and Contract: The Case of Strikes," *Journal of Political Economy*, October 1980, *88*, 867–86.

Riddell, W. Craig, "The Effects of Strikes and Strike Length on Negotiated Wage Settlements," Department of Economics Discussion Paper 80-2, University of British Columbia, 1980.

Schnell, John and Gramm, Cynthia, "Crossing the Picket Line: An Analysis of Player Choice During the 1987 Football Strike," paper presented at the Industrial Research Association Annual Meeting, December 1989.

Sobel, Joel and Takahashi, Ichiro, "A Multistage Model of Bargaining," *Review of Economic Studies*, July 1983, *50*, 411–26.

Tracy, Joseph, "An Investigation into the Determinants of U.S. Strike Activity," *American Economic Review*, September 1986, *76*, 423–36.

______, "An Empirical Test of an Asymmetric Information Model of Strikes," *Journal of Labor Economics*, April 1987, *5*, 149–73.

Vroman, Susan, "A Longitudinal Analysis of Strike Activity in U.S. Manfacturing: 1957–1984," *American Economic Review*, September 1989, *79*, 816–26.

Negotiator Behavior and the Occurrence of Disputes

By Orley Ashenfelter and Janet Currie*

It is generally recognized that gains from cooperation are sometimes lost because negotiating parties fail to reach agreement. Table 1 shows dispute rates that have been reported in a wide variety of negotiating situations. The estimates suggest that between 10 and 49 percent of negotiations end in a dispute. Such negotiation failures are *ex post* inefficient and have long defied economists' attempts to explain them. After all, if the parties know the probability distribution of the outcomes that may be forced upon them by an arbitrator, judge, or jury, why do they not reach an agreement that avoids costly conflict? In this paper, we discuss the recent empirical research on the causes of disputes.

Our focus is primarily on the analysis of disputes that are resolved by arbitration systems. In practice, nearly all of the empirical research on the determinants of disputes (apart from that devoted to the study of strikes in labor negotiations) has been devoted to the study of arbitration systems. Since 1960, compulsory arbitration has become an increasingly common means of resolving contract disputes in the American public sector. Half of the states now have arbitration statutes covering some group of public employees. Arbitration is also widely used to settle labor disputes in Canada and Europe, and as an alternative to the court system in the United States. In short, arbitration systems have become a rich and varied source of data for empirical tests of alternative theories of the causes of disputes.

Arbitration systems are highly structured. The parties first agree on an arbitrator.[1] They then present their cases, and the arbitrator imposes a settlement. Under conventional arbitration, the arbitrator can impose any settlement, while under final-offer arbitration, the arbitrator must choose the position of one of the parties. Tri-offer arbitration, a variant of final-offer arbitration in which the arbitrator may also choose a fact-finder's position, has been used in Iowa since 1974.

The study of these more structured approaches to dispute resolution offers additional insights into the causes of negotiation failures. We focus on four factors that are

TABLE 1—REPORTED DISPUTE RATES[a]

Author	Data	Rate
Collective Bargaining: Strikes		
Card	Canadian private sector contracts	22
McConnell	U.S. private sector contracts	17
Currie-McConnell	Canadian public sector contracts	13
Collective Bargaining: Arbitrations		
Currie-McConnell	Canadian public sector contracts	32
Currie	British Columbia teachers	33
Ashenfelter-Bloom (1984)	New Jersey police	30–49
Ashenfelter et al.	Arbitration experiments	28–43
Boden	Worker's compensation	43
Other Types of Negotiations		
Roth-Ochs	Two-person bargaining games	15
Mnookin et al.[b]	California child custody	22
White[b]	Medical malpractice	11

[a]Rates are shown in percent.
[b]Percent of cases that go to court.

---

*Professor, Department of Economics, Princeton University, and Assistant Professor, Department of Economics, UCLA, Los Angeles, CA 90024, respectively. We thank Duncan Thomas for helpful comments on an earlier draft.

[1] David Bloom and Christopher Cavanagh (1986) offer an insightful empirical analysis of the way in which arbitrators are chosen. Ashenfelter (1987) explains why the emerging evidence implies that arbitrators operate within a process that makes them statistically exchangeable.

thought to be important determinants of dispute rates both under arbitration systems and in the courts. They are 1) divergent beliefs about the expected outcome if the dispute must be arbitrated; 2) principal-agent problems where it is in the interests of a principal, but not his or her agent, to proceed to arbitration; 3) attitudes towards the uncertainty associated with going to arbitration; and 4) dispute costs.

I. Divergent Expectations

Vincent Crawford (1979) points out that if both parties to a negotiation knew the arbitrator's preferred outcome, they would never use the procedure: they would simply settle on that outcome and avoid the arbitral costs. Thus, in order for the procedure to be used at all, there must be some uncertainty associated with the arbitral outcome.[2] Moreover, as Ashenfelter (1987) points out, it is very unlikely that the parties would ever voluntarily agree to submit their dispute to a deterministic system since this is tantamount to agreeing on a settlement.

In a pioneering early paper, Henry Farber and Harry Katz (1979) show that if the parties were risk averse and had identical expectations about the uncertain arbitral settlement, then there would be a range of settlements that both parties would prefer to incurring the cost and the uncertainty associated with arbitration. It follows that the parties may go to arbitration solely because they have differing beliefs about the expected outcome of arbitration. Intuitively, if each party believes that it will "win" in the arbitration hearing, then the incentive to reach a negotiated settlement is diminished.

The role of divergent expectations is similar to the role of different beliefs about relative delay costs in the "war of attrition" model of strikes discussed by John Kennan and Robert Wilson (1988). In both cases, the parties know the size of the pie to be divided, but they have differing beliefs about the likely winner in the event of a dispute. This analogy suggests that divergent expectations might reflect private information. Farber and Max Bazerman (1989) cite the psychology literature and argue that divergent expectations might also reflect a human tendency to overestimate the probability of success in any uncertain situation. A question they do not address is, to what extent learning would cause expectations to converge in the course of repeated negotiations.

II. Principal-Agent Problems

A recurrent theme in the Industrial Relations literature about arbitration is that union negotiators may have incentives which are different from those of the rank and file. Brian McCall (1990) appears to be the first author to formalize this idea, and he does so in the context of interest arbitration. In his model, the negotiator can procure a better settlement for the union by expending more effort. The settlement is also affected by the state of the world, which is observable to the negotiator but not to the union rank and file. Hence, the rank and file cannot distinguish between a bad state caused by the negotiators shirking and one due to a bad state of the world. McCall shows that requiring the negotiator to use arbitration in the bad state will solve the incentive compatibility problem.

Although the model's principal prediction is that arbitration will be used only in the bad state of the world, the available evidence suggests that there is no difference in wages between arbitrated and negotiated settlements within jurisdictions (compare Currie, 1989). However, McCall does point out that if there is also a principal-agent problem on the management side, then there will not be any simple relationship between arbitration and wages.

Ashenfelter and David Bloom (1989) have also set out a model where disputes may arise when the agent's role is to manipulate the outcome, as do lawyers in arbitration and legal proceedings. In such situations, the parties may well face incentives similar to those in a prisoner's dilemma payoff structure. It is a dominant strategy for each party to engage the agent, who in turn causes the parties to engage in a dispute in order to collect a larger fee. It is well known that it pays for the parties to write a contract that specifies cooperation in this situation, but it is also known that such contracts may be difficult to enforce. Although Ashenfelter and Bloom present evidence that indicates that several arbitration systems appear to contain a prisoner's dilemma payoff structure, it is not possible from this evidence alone to determine the extent to which the observed number of disputes is greater than would be the case in the absence of any agency problem.

---

[2] The same observation can be made about the court system. If the outcome of a trial were known with certainty before the trial began, we would expect litigants to settle out of court.

III. Attitudes Towards Risk

If the parties to a negotiation are risk averse, the degree of uncertainty associated with arbitral outcomes will also affect dispute rates. Other things being equal, an increase in uncertainty should reduce dispute rates. Ashenfelter et al. (1989) report results of an experiment designed to test this hypothesis:

Each pair of subjects bargained over a sum of money twenty times. During the first ten rounds of bargaining, they each received a zero payoff if they were unable to agree within a fixed time. During the next ten rounds, if they could not agree within the fixed time the division was decided by "arbitration." The arbitral decision was modeled as a random number drawn from a normal distribution. Thus the uncertainty associated with arbitration could be varied by changing the variance of the distribution of arbitral decisions. Table 2 shows dispute rates associated with different levels of uncertainty. The first column shows that as the variance of the distribution of arbitral outcomes was increased, dispute rates fell.

TABLE 2—EXPERIMENTAL DISPUTE RATES[a]

	(1)	(2)
No Arbitration Controls		
($\sigma = 50$)	–	8/200
Conventional Arbitration		
Low Variance ($\sigma = 12.5$)	119/289	40/290
Medium Variance ($\sigma = 25$)	79/250	26/250
High Variance ($\sigma = 50$)	71/250	24/250
Final-Offer and Tri-Offer (Combined)		
($\sigma = 50$)	216/537	66/540

Notes: Parties bargained over a number between 100 and 500. Arbitral preferences were distributed normally with mean 350 and a standard deviation equal to σ. No arbitration controls bargained all 20 rounds without arbitration.

[a](Number of Disputes)/(Number of Negotiations) in the Ashenfelter et al. experiments with arbitration (col. 1) and without arbitration (col. 2).

It is likely that attitudes to risk vary across bargaining pairs. For example, Ashenfelter and Bloom (1984) offer evidence that some parties to arbitration hearings are more risk averse than others. In an examination of final-offer arbitration of New Jersey police contracts, they found that unions made offers that were more conservative than those of their employers—in the sense that they were closer to a measure of the mean arbitrator's preferred award. Several authors have shown that there is a great deal of unexplained variability in dispute rates which is associated with individual bargaining pairs (compare Currie, 1989, Ashenfelter et al., and Melvyn Reder and George Neumann, 1980). Differing attitudes towards risk might help explain this variation.

Finally, Farber and Katz suggest that parties who use arbitration learn about arbitrator preferences and reduce their subjective evaluation of the degree of uncertainty associated with the procedure. Hence, risk aversion offers a potential explanation for state dependence in the occurrence of arbitrations Currie (1989) shows that parties who used arbitration in the last negotiation are more likely to use arbitration in the current round of negotiations, but that there is no evidence of a more enduring "narcotic effect" of arbitration.

IV. Dispute Costs

Bloom (1981) suggests that, in some cases, the cost of continuing negotiations may be high enough that it will be less costly to use arbitration. Whatever the cause of bargaining failures, any economic theory of disputes predicts an inverse relationship between dispute rates and the cost of disputes. In a bargaining experiment with 171 sets of negotiations, Farber et al. (forthcoming) show that higher direct costs of arbitration are associated with a decrease in dispute rates.

Little is known about the relative costs of disputes under different dispute resolution procedures, but the available evidence suggests that arbitration hearings are much less costly than strikes.[3] This observation is no doubt the explanation for the fact that a regime in which disputes are resolved by compulsory arbitration is associated with far more disputes than a regime in which disputes result in a strike: it is simply that disputes are more common when they are less costly. Some evidence of this "chilling effect" is shown in Table 2: a comparison of columns 1 and 2 shows that the experimental dispute rates were about three times higher in the rounds with arbitration than in the rounds without.

As a second example, Currie and McConnell study 3985 contracts covering Canadian provincial and municipal level employees between 1965 and 1987, and show that as reported in Table 1, 32 percent of negotiations subject to compulsory arbitration ended up in disputes compared to 13 percent for contracts negotiated under the right to strike. Controlling for occupation and jurisdiction does little to narrow the raw difference in dispute rates. Currie and McConnell show that, even after taking into account the increase in dispute rates associated with arbitration, the total cost of contract disputes as measured by foregone wages is lower under compulsory arbitration than under the right to strike.

V. Conclusion

It is possible that 1) divergent expectations, 2) principal-agent problems, 3) attitudes towards risk, and 4) dispute costs may all play a role in negotiation failures. Our discussion indicates that the role of these factors in the determination of dispute rates is only beginning to be understood. It is our hope that the insights and empirical methods gained from the study of arbitration systems can be used to develop a more general theory of negotiation failures that will allow us to address an important set of questions that have long resisted economic analysis.

---

[3] Currie and Sheena McConnell (1989) summarize the available estimates and conclude that the cost of an arbitration hearing would be most unlikely to exceed $50,000 Canadian.

REFERENCES

Ashenfelter, Orley, "Arbitration and the Negotiation Process," *American Economic Review Proceedings*, May 1987, *77*, 342–46.

Ashenfelter, Orley, and Bloom, David, "Models of Arbitrator Behavior: Theory and Evidence," *American Economic Review*, March 1984, *74*, 111–25.

______ **and** ______, "Lawyers as Agents of the Devil in a Prisoner's Dilemma Game," mimeo., Princeton University Industrial Relations Section, November 1989.

Ashenfelter, Orley et al., "An Experimental Comparison of Alternative Arbitration Systems," mimeo., UCLA, December 1989.

Bloom, David, "Is Arbitration Really Compatible with Bargaining?," *Industrial Relations*, Fall 1981, 20, 233–44.

______ **and Cavanagh, Christopher,** "An Analysis of the Selection of Arbitrators," *American Economic Review*, June 1986, *76*, 408–22.

Boden, Leslie, "Dispute Resolution in Worker's Compensation: The Role of Adversarial Experts," mimeo., Boston University, October 1989.

Card, David, "Strikes and Wages: A Test of a Signalling Model," mimeo., Princeton University, January 1988.

Crawford, Vincent, "On Compulsory Arbitration Schemes," *Journal of Political Economy*, February 1979, *87*, 131–60.

Currie, Janet, "Who Uses Interest Arbitration? The Case of British Columbia's Teachers 1947–1981," *Industrial and Labor Relations Review*, April 1989, *42*, 363–79.

______ **and McConnell, Sheena,** "Strikes and Arbitration in the Public Sector: Can Legislation Reduce Dispute Costs?," mimeo., UCLA, August, 1989.

Farber, Henry and Bazerman, Max, "Divergent Expectations as a Cause of Disagreement in Bargaining: Evidence from a Comparison of Arbitration Schemes," *Quarterly Journal of Economics*, February 1989, *54*, 99–120.

______ and Katz, Harry, "Interest Arbitration, Outcomes and the Incentive to Bargain," *Industrial and Labor Relations Review*, October 1979, *33*, 55–63.

______, Neale, Margaret and Bazerman, Max, "The Impact of Risk Aversion and Costs of Arbitration on Resolution Rates and Negotiated Outcomes," *Industrial Relations*, forthcoming.

Kennan, John and Wilson, Robert, "Strategic Bargaining Models and Interpretation of Strike Data," mimeo., University of Iowa, November 1988.

McCall, Brian, "Interest Arbitration and the Incentive to Bargain: A Principal-Agent Approach," *Journal of Conflict Resolution*, March 1990, *34*, 151–167.

McConnell, Sheena, "Strikes, Wages and Private Information," *American Economic Review*, September 1989, *79*, 801–15.

Mnookin, Robert et al., "Private Ordering Revisited: What Custodial Arrangements are Parents Negotiating?," mimeo., Stanford University, August 1989.

Reder, Melvyn and Neumann, George, "Conflict and Contract: The Case of Strikes," *Journal of Political Economy*, October 1980, *88*, 867–86.

Roth, Alvin and Ochs, Jack, "An Experimental Study of Sequential Bargaining," mimeo., University of Pittsburgh, February 1988.

White, Michelle, "Dispute Prevention and Dispute Resolution in Medical Malpractice," mimeo., University of Michigan, November 1989.

THE INDUSTRIAL ORGANIZATION OF MARKETS WITH DIFFERENTIATED PRODUCTS[†]

The Introduction of New Products

By DEBRA J. ARON AND EDWARD P. LAZEAR*

How do firms decide whether to enter new markets? There is abundant economic literature on *R&D* and patent races,[1] but little that actually addresses the order of entry or type of firm that moves into new markets.

We argue that new markets are likely to be opened up by new entrants, and firms that are less dominant in existing markets. There are a few reasons for this. First, firms that are behind in one market enjoy a relative benefit from high-variance strategies. This is akin to the practice by trailing football teams of throwing long passes in the last few minutes of the game. A low-variance strategy is almost certain to result in defeat. What the team seeks is a fat-tail strategy that gives some significant probability of winning, even if the strategy implies fewer expected points scored. Second, cannibalization is something that affects incumbents and new entrants asymmetrically. A new firm introduces a product when the revenues from doing so exceed the costs. An incumbent's calculation must include in costs lost revenues from sales of the existing product when the new one is introduced. Third, there may be diseconomies of scope that affect incumbents differently from new entrants. If an incumbent's cost of producing its current product is adversely affected by the introduction of a new one, then incumbents will be less likely to innovate than new firms, or firms that do not produce in a currently related market.

The timing of product introduction is central to our story. While incumbents may be reluctant to open up new markets, they can be induced to follow into the new product line. (Some other studies that have examined the timing of innovation are Lazear, 1974, and Andrei Shleifer, 1986.) It may not pay to lead, but given that another firm produces the new product, the incumbent may find that it is better to follow than to remain out. Uncertainty about entry is not necessary to achieve this result. The incumbent may not want to lead because it alters the rival's decision in a way that is detrimental to the incumbent. Instead, the incumbent may prefer to follow because the costs of entry, including lost sales, need not be borne until another firm establishes a foothold in the new product.

Despite the incumbent's unwillingness to lead, products tend to be introduced nearly simultaneously by multiple firms. For example, *Newsweek* and *Time* often have virtually identical cover stories. This is to be expected when there is a major news event, but the model described below rationalizes less obvious cases where AIDS or Bruce Springsteen are featured by both magazines in the same week.

These issues imply that entry decisions by new firms and counters by incumbents are not always efficient. Because entrants do not take into account the effect on fixed costs and diseconomies of scope borne by incumbents, they enter too often. Given entry by outsiders, the incumbent's decision to follow into the new market is efficient, but only relative to staying in the old product while

---

[†] *Discussants*: Yukiko Hirao, University of Pennsylvania; Pablo Spiller, University of Illinois-Urbana; Jonathan Hamilton, University of Florida.

*Managerial Economics and Decision Sciences Department, Northwestern University, Evanston, IL 60208; Graduate School of Business, University of Chicago, Chicago, IL 60637 and Hoover Institution, Stanford, CA 94305, respectively. Financial support from the National Science Foundation is gratefully acknowledged.

[1] See, for example, Kenneth Judd (1985), Glenn Loury (1979), and Jennifer Reinganum (1981, 1982, 1985).

the entrant produces the new. The reason is that, given that entry has occurred, the only costs borne are by the incumbent. The rival's costs are now sunk so the incumbent's *ex post* decision is efficient.

We have examined the issues in detail in two earlier papers.[2] In this paper, the model used in the stochastic version of the analysis is sketched. Results from the deterministic version, that incorporates cannibalization and diseconomies of scope, will be referenced.

I. A Sketch of the Stochastic Model

The model makes two points. First, firms that are behind seek variance and the introduction of new products provides the desired variance. Second, the industry's dominant firm is likely to follow when the fortunes of one firm are highly correlated with the fortunes of other firms in the new market. Thus, the trailing firm leads and the old industry leader follows. A somewhat whimsical example from the world of yacht racing illustrates the point.

In 1983, the United States lost the America's Cup competition for the first time in the event's history. Experts commented that the United States lost the cup in part because the skipper violated a fundamental rule of yacht racing. The typical situation was as follows. The U.S. yacht had a substantial lead over the second-place Australian team. With both boats sailing the same course of direction, the Australian team would almost surely lose if it remained on the same course. Thus, the best strategy for the second-place boat was to tack. To tack is to veer off at an angle from the current course of direction and take a saw-toothed path to the finish. One does this in the hopes of benefiting from an unexpected shift in the wind that would make the longer, indirect course the faster one. The rule of racing is that the leader should follow the strategy of his opponent in selecting sailing direction. Even though the initial path may have a shorter expected finish time, the probability of winning is maximized by switching if the second-place opponent switches, because the leader maintains his relative advantage over the challenger. If the wind shifts favorably for the challenger, the leader benefits as well. If it does not, the leader is still likely to maintain his lead. The key is that randomness in wind associated with choosing the new path is correlated across competitors.

The yacht race differs from product markets in two respects. First, in the race, only rank-order position matters, as is the case in labor market tournaments (Lazear and Sherwin Rosen, 1981). In markets, profits depend on the size of the lead as well. Second, firms can produce in both markets simultaneously, albeit at some additional cost. The latter point is not essential, since we can think of strategy 1 as being in the old product and strategy 2 as being in both products. The first point is somewhat more important.

To make the analysis more formal, consider two firms, A and B. Each produces its version of some product X. The costs of production are the same for both firms, but the products are not identical from the perspective of the consumers. Indeed, even if firm B charged price equal to marginal cost for its product, all consumers would prefer A's version over a large range of prices exceeding marginal cost (this will be specified more clearly presently). Thus, we consider a market in which one firm has gained advantage in a product market by virtue of marketing a product that is superior in the consumers' eyes.

The attributes of a firm's product are idiosyncratic and cannot be imitated by rival firms. Each firm has a unique set of attributes and the expected value of each firm's product may be the same, but *ex post*, after consumers reveal their preferences, one product is found to be superior.

A firm can, at any time, choose to enter into production of a different product, Y, that is a substitute for X to some degree. By entering another product, the firm must divert resources from production of X. We consider the limiting case in which the firm must exit X entirely in order to produce Y. In this version, for simplicity, there is an extreme diseconomy of scope.

---

[2] See our 1987 and 1989 papers. This paper is closer to the 1987 paper. The model there is stochastic and firms are permitted to compete in only one industry at a time. In this paper, the stochastic model is discussed, but reference to the 1989 results, where multiple products may be produced, is made.

Demand for Y is unknown *ex ante*. It is a new product, and could "hit" or "miss." One characterization implies that if Y hits, it will, over some range of prices exceeding marginal cost, supplant demand for X entirely; if it misses, and X is still on the market, no one will purchase Y. However, if X is unavailable, there will be some demand for Y even when it misses. Both firms A and B could produce a version of Y and, again, these would not be identical. Since consumers see the two versions as imperfect substitutes, either, both, or neither version could be a hit. To the extent that Y_A and Y_B are similar products, however, the probability that Y_A hits will be correlated with the probability that Y_B hits. But correlation is unlikely to be perfect. Since A has proven to produce a superior product in the X market, one might suspect that A would have a higher probability of success in Y than does B. The idiosyncratic feature of firm A generating the success of X_A is likely to carry over into the Y market.

In this scenario, firms make decisions about which market they choose to operate in. Their decisions will, of course, depend on the payoffs in each market, which are endogenously determined by the consumers' preferences, the firms' costs, and the strategic interaction between the firms given their market choice. Assume that firms behave as Bertrand oligopolists, and parameterize consumers' preferences as follows:

$$U = f\left[\left[(\alpha X_A + X_B) + (\gamma Y_A + \beta Y_B)\right], m\right], \tag{1}$$

where α is known, $\alpha > 1$, γ and β are unknown, and m is a vector of other goods. In this utility function, X_A, X_B, Y_A, and Y_B are perfect substitutes in proportions determined by the parameters α, β, and γ. This is an implication that captures the essence of the general case in which the products are substitutes to some degree, but not necessarily perfect substitutes.[3]

TABLE 1

		Firm A		
		γ_H	γ_L	
Firm B	β_H	$P - z$	z	P
	β_L	z	$1 - P - z$	$1 - P$
		P	$1 - P$	1

II. Assumptions

We specify the distribution of γ and β as $\gamma = \{\gamma_L, \gamma_H\}$, $\beta = \{\beta_L, \beta_H\}$, where L denotes low and H denotes high. For simplicity, assume that $\text{Prob}(\gamma_H) = \text{Prob}(\beta_H) = p$, where β, γ may be correlated. Let $z \equiv \text{Prob}(\gamma_H, \beta_L)$. Then the joint distribution of γ, β is shown as Table 1. If both A and B are in Y, then firm B "wins" only z of the time, when $\beta = \beta_H$ and $\gamma = \gamma_L$. The following assumptions make this true.

A1: $\gamma_H > \beta_H$, $\gamma_L > \beta_L > 0$;

A2: $\gamma_L < \alpha$, $\beta_H > \alpha$; A3: $\alpha > 1$;

A4: $MC_X^A = MC_Y^A = MC_X^B = MC_Y^B = 1$.

A1 means that if both firms' version of Y were hits, or both were misses, then consumers would prefer Y_A to Y_B at equal prices. This corresponds to the earlier argument that if A is successful in dominating B in the X market, the same firm is more likely to dominate in the Y market as well.

A2 is needed so that it may pay for the firm that is behind to switch to Y; specifically, $\beta_H > \alpha$ means that if A stays in X and B switches to Y and Y hits, then B wins. It also implies that the firm in the lead does not automatically switch to production of Y; specifically, $\alpha > \gamma_L$ means that if Y misses, A would have been better off continuing to produce X.

A3 assures that consumers prefer X_A to X_B, and A4 is merely a convenience. Since we want to focus on uncertain product demand, we assume that costs of production are independent of product choice and nature and normalize marginal cost to 1.

---

[3] As contrasted with Jeremy Bulow et al. (1985), the goods are substitutes on the demand side, rather than in production.

Once a firm decides to move to the Y market, it takes some time to effect the change; the new product must be developed, personnel must be trained, contracts must be written with new suppliers and terminated with old ones, plants must be restructured. This is publicly observed behavior. The decision to commit to Y is an announcement, but production does not begin until after some delay, T. If firm i decides at time t to enter Y, firm j could follow immediately or wait until firm i produces. By waiting, j discovers whether firm i's product is a hit or a miss before committing to Y. The cost of waiting is that if it subsequently decides to produce Y, firm j must then wait for time period T until it is ready to begin production of Y. The firm could also wait some intermediate amount of time before announcing its commitment to entering Y, but it will always be optimal to either commit immediately in response to the rival firm, or to wait until it actually produces in order to learn whether the product is successful.

At any moment in time, each firm will choose its action based on whether or not its opponent has moved, and if so, what, if anything, has been learned. Once a firm's opponent has committed itself to product Y, a firm faces a simple, nonstrategic optimization problem. It can choose to commit to product Y immediately, or it can choose to wait. We impose subgame perfect behavior on the firms, so no firm can commit *ex ante* to do something that would be suboptimal *ex post*.

The gain to waiting is that if the opponent's product is a failure, firm j can choose not to enter Y. On the other hand, if i waits and Y_j is a success, then j suffers a loss for the entire period required for j to develop its version of Y. Thus, depending on how long is required to develop a new product, a firm may choose to follow its opponent into a new market without waiting to see if the product is successfully produced *ex ante*. We are particularly interested in the equilibrium in which the firm that is losing in market X enters market Y, and the leader in the X market follows the rival into Y.

III. Results

Define $\pi_j(R,S)$ as firm j's expected profits when firm A produces product R and firm B produces product S. The worst situation for B is (X, X), which yields $\pi_B(X, X) = 0$.

The most important result is that firm B will lead into Y and firm A will follow under some easy-to-meet conditions. Specifically, let

$$(2) \qquad \pi_B(Y,Y) < \pi_B(Y,X);$$

$$(3) \qquad \pi_A(X,X) > \pi_A(Y,X).$$

Firm B leads into Y for the following reason. If B does not lead into Y and A stays in X, firm B is a sure loser. The only reason that B would not lead into Y is that B thinks that A will lead if B waits. But if A leads, A knows that B will not follow because of (2). Firm B prefers to remain in (or move into) X if A vacates. Further, (3) implies that firm A prefers (X, X) to (Y, X) so A would not lead if it were convinced that B would stay in X. By refusing to lead, A forces B to lead because B's refusal to do so results in $\pi_B(X, X)$, the worst of all worlds from B's point of view.

Further, if additionally,

$$(4) \qquad \pi_A(Y,Y) > \pi_A(X,Y)$$

and set-up time is sufficiently long, then firm A follows immediately into Y. Condition (4) merely says given that firm B has led into Y, firm A prefers following to remaining in X. If set-up time is sufficiently long, A moves immediately because the gains from having the product on line sooner rather than later outweigh the advantages of seeing how B fares before following. Thus, both *Time* and *Newsweek* feature Bruce Springsteen in the same week.

What underlies firm B's move to Y is that it has "lost" in the X market. It is in B's interest to take a risky strategy. If B moves first, it knows that A will follow when (4) holds. A positive correlation in outcomes induces firm A to follow into Y, but even if both compete in the Y market, firm B still has some positive probability of winning. Like the trailing team during the last few minutes of a football game, the firm that is behind tends to compete in the high variance market. Further, the weaker firm in the current market "draws in" the leader to the new market. The weaker firm enters the new market first, and the leader follows immediately. We will refer to this as the "follow" equilibrium.

One other key point relates to the likelihood that firm A follows firm B into Y. The set of parameters over which the market leader follows the maverick (immediately) into the new product line (i.e., the follow equilibrium) increases in the covariance between the probability of each new product being a hit, given variance. If A's fortune in Y is highly correlated with B's, then A is more likely to follow, even before outcomes are revealed. A knows that a hit by B will leave A in bad shape. When A's and B's outcomes in Y are highly correlated, then A knows further that a hit by B most likely connotes a hit by A. Firm A can counter B's move by following B into the new product. Thus, if *Newsweek*'s ability to do well with a story on Springsteen signals that *Time* is also likely to do well with the same story, then *Time* will follow *Newsweek*'s lead. The ability to respond during the same week requires that one magazine can obtain information on what the other is doing with enough time left to copy. In the model, it is the announcement time that is relevant.

Finally, given that firm B leads, firm A is less likely to follow, the larger was A's lead (say, in terms of profit differences) in the old market. Using the football analogy, the leading team is unlikely to match the trailing team's high-variance strategy when the lead is virtually insurmountable. IBM was not about to drop its mainframe 370 model to follow Cray into the realm of supercomputers—IBM's initial lead was too great. Diseconomies of scope are key here. Otherwise, IBM could produce both.

As the number of firms increases, following becomes more likely. The reason is that it may be viable to remain in X when only firm B exists because B may not hit. But if there are a large number of firms in Y, the chances that at least one will hit increase. In the limit, firm A is a certain loser if it remains in X—it is thereby forced to adopt a higher variance strategy as well. The folklore in the personal computing industry is that IBM shed some of its conservative ways and became more aggressive when it entered the PC market. The existence of a large number of viable PC competitors could explain the change.

IV. Efficiency

Does firm B enter Y only when it is efficient to do so? Does firm A follow when and only when it is efficient? The general answer is no, due to fixed costs and diseconomies of scope (see our 1989 paper). Consider less extreme diseconomies of scope so that A continues to produce X even after it produces Y. When it enters Y, firm B cares only about its own expected profits, ignoring any social costs it might impose on A. Those costs take two forms. First, there is a duplication of fixed costs, which A bears to enter Y. Second, A's entry into Y, although profitable, makes X production less efficient.

For example, note that it is possible that after one period's production, firm A is so efficient that it can drive firm B out of Y altogether. This gives A an incentive to enter Y to recapture sales. The preferred solution may well have been for B to stay out altogether, and for A to produce only X, avoiding the costs of entering Y and the reduction in efficiency of X production.

V. Conclusion

Products tend to be introduced by more than one firm simultaneously. Part of this is attributable to technology. More than one firm may have access to an innovation that makes a new product viable. But, beyond this, firms move together because of their strategic involvement with one another. Further, firms that are behind in old markets are likely to lead into new ones. Industry giants prefer conservative strategies, whereas their fringe rivals seek variance that gives the change of victory to an otherwise beaten firm. The maverick can induce the incumbent to follow into the new line because firms' successes in the new market are correlated. As the number of rivals increases, the likelihood that the incumbent follows increases as well. Finally, because entrants ignore costs imposed on the incumbents, the outcome is not always efficient.

REFERENCES

Aron, Debra J. and Lazear, Edward P., "Competition, Relativism, and Market Choice," Working Papers in Economics E-87-56, Hoover Institution, December 1987.

______ and ______, "Competition and Market Choice," University of Chicago, October 1989.

Bulow, Jeremy I., Geanakoplos, John D. and Klemperer, Paul D., "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, June 1985, *93*, 488–511.

Judd, Kenneth, "Closed Loop Equilibrium in a Multistage Innovation Race," Center for Mathematical Studies in Economics and Management Sciences, Northwestern University, Discussion Paper No. 647, February 1985.

Lazear, Edward P., "The Timing of Technical Change: An Analysis of Cyclical Variations in Technology Production," unpublished doctoral dissertation, Harvard University, 1974.

______ and Rosen, Sherwin, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, October 1981, *89*, 841–64.

Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *89*, 395–410.

Reinganum, Jennifer F., "Dynamic Games of Innovation," *Journal of Economic Theory*, August 1981, *25*, 21–41.

______, "A Dynamic Game of R&D: Patent Protection and Competitive Behavior," *Econometrica*, May 1982, *50*, 671–88.

______, "Innovation and Industry Evolution," *Quarterly Journal of Economics*, February 1985, *100*, 81–99.

Shleifer, Andrei, "Implementation Cycles," *Journal of Political Economy*, December 1986, *94*, 1163–90.

Pricing Behavior and Vertical Contracts in Retail Markets

By ANDREA SHEPARD*

Recent developments in the economic theory of multifirm markets and firm organization have motivated several empirical studies of retail gasoline markets. A growing body of theoretical work suggests that the behavior of firms in markets once viewed as "workably competitive" can diverge significantly from that characteristic of competitive markets. Static models of free-entry markets where firms are horizontally differentiated suggest that the equilibrium price structure will reflect local market power. Dynamic oligopoly models suggest that repeated interaction can sustain (tacitly) collusive outcomes in multifirm industries. Because gasoline stations are strikingly simple firms, and data are available on wholesale and retail gasoline prices, this market has proven to be a useful arena for testing pricing predictions from these models. These tests are discussed in Section I.

Another growth area in industrial organization is research on the nature of vertical contracts. This work suggests that the optimal contract between a manufacturer and a retailer will be sensitive to the nature of downstream competition and the available control technology. Imperfect competition downstream can lead to double marginalization, and imperfect monitoring can allow shirking or free riding. The manufacturer can be expected to choose contractual forms that minimize these effects. Section II discusses efforts to assess the effect of competition and monitoring on the refiner's choice of contractual form for retail outlets.

While examining contractual form is interesting in itself, policy initiatives provide additional impetus for examining contracts in gasoline retailing. Refiners have long been a favorite target of antitrust enforcement, and complaints regarding restraints imposed on retailers have been viewed sympathetically by the courts. Court decisions have made unlawful exclusive dealing clauses for gasoline and automotive products as well as the various arrangements through which a refiner can control the retail price at franchise outlets. In the last decade, federal and state governments have enacted legislation further restricting the nature of the contract between refiners and the retail outlets for their gasoline. The growth in these restrictions naturally leads to positive questions about the nature and effect of contractual relationships between refiners and dealers.

I. Pricing Behavior

One view of gasoline retailing is that stations compete in price and behave approximately Bertrand, that is, prices are defined by cost. This view is supported by the observation that stations post prices, competitors can respond immediately to price changes, branded gasoline of a specified octane level is a homogeneous product, and there are many stations. Contrary to this view, there is a growing body of empirical work that suggests these markets depart from the Bertrand paradigm in significant ways.

Price behavior in an isolated market with thirteen stations has been studied by Margaret Slade (1986, 1987) using daily observations on price in a period characterized by price instability induced by demand shocks. Slade (1986) estimates a conjectural variations model and finds that price-taking behavior can be rejected. She also estimates station demand equations and finds that both own-price and cross-price demand elasticities show substantial variation across stations, even though these stations are located along the same heavily traveled commercial strip. Using these data in a dynamic model to estimate the slopes of linear intertemporal

---

*Economics Department, MIT, Cambridge, MA 02139. This research is supported by the Center for Energy Policy Research, MIT.

TABLE 1—RETAIL MARGINS[a]

	No. of Stations	Average Margin	Standard Deviation
All Gasoline:			
Regular Leaded	1232	28.4 (0.21)	7.2
Regular Unleaded	1512	35.6 (0.22)	8.6
Premium Unleaded	1441	40.0 (0.24)	9.0
Self-Service Gasoline:			
Regular Leaded	472	24.5 (0.21)	4.5
Regular Unleaded	518	30.7 (0.24)	5.3
Premium Unleaded	500	35.1 (0.27)	6.1
Full-Service Gasoline:			
Regular Leaded	969	33.3 (0.29)	9.2
Regular Unleaded	1230	40.7 (0.29)	10.2
Premium Unleaded	1174	44.8 (0.29)	10.1

[a] Margins: in cents. Standard errors are shown in parentheses.

reaction functions, Slade (1987) finds support for the hypothesis that the stations play supergame strategies. In particular, she finds that the estimated reaction function slopes are significantly different from single-period best replies, and result in higher estimated profits. These results suggest that brand and location provide sufficient local market power to allow firms to price above marginal cost, and that station operators are sufficiently sophisticated to exploit repeated play in order to increase profits.

Gasoline is sold in a variety of grades (regular leaded, regular unleaded, etc.) and can be sold self-service or full-service. In a perfectly competitive market, each grade and service level would be sold at marginal cost. In a free-entry market in which gasoline revenues must cover a fixed cost, prices may be above marginal cost, but will reflect fully allocated fixed cost. A cursory examination of data on retail margins suggests that the price structure is more complex than these cost-based stories would suggest.

Table 1 presents data on retail margins (retail prices minus wholesale prices) by grade and service level. The retail price data are from a cross-sectional census of gasoline stations in eastern Massachusetts in early 1987, and include observations at over 1500 stations. The wholesale prices are the average weekly FOB prices for Boston.[1] The first column reports the number of stations carrying the specified grade. There are, for example, 1232 stations selling regular leaded gasoline. The retail margins reported in the top three rows use the lowest price at the stations for the specified grade. At a station selling both full- and self-service gasoline, this will be the self-service price. The six rows at the bottom separate full-service and self-service prices "Mini-service," in which an attendant pumps the gas but will provide no other service, is treated as self-service.

The retail margin is higher for full-service than self-service: the difference in means is 9 to 10 cents per gallon for each grade.[2] Since full-service sales involve a higher labor input than self-service, it would be surprising if this were not the case. More interesting is the difference in margins by grade; the margin on premium unleaded is over 40 percent higher than the margin on regular leaded. This difference might be cost-based if gasoline were priced to cover grade-specific fixed cost, and the sales volume of premium unleaded were much smaller than the sales volume of regular leaded. Premium unleaded sales, however, were almost 50 percent higher than regular leaded sales for the state as a whole in late 1986 (U.S. Department of Energy, *Petroleum Marketing Monthly*, 1987). The margin differences might also be cost-based if buyers of premium unleaded have a higher demand for service. If, for example, a premium gasoline purchaser is more likely to have the station attendant check her oil. But, since service comes only with the purchase of full-service gas, this would imply greater margin difference for full-service purchases. As shown in Table 1, however, the difference is approximately the same for full-service and self-service.

---

[1] Although prices at any given station in the census were reported only once, data collection extended over 13 weeks. The wholesale price used to construct retail margins is the price for the week in which the retail price was observed.

[2] All differences in means or variances mentioned in the text satisfy statistical significance at the 0.001 level.

Another possible explanation for this dispersion in retail margins is that stations have sufficient local market power to price discriminate. Consumers of leaded and unleaded gas are segmented by law and by the technology embodied in automobiles. Premium vs. regular unleaded purchasers may also be segmented by technology (some producers of high performance cars recommend premium fuel), but the segmentation seems less clear. Consumers also might be separable based on willingness to pay for quality. If unleaded premium and regular gasolines are perceived as vertically differentiated goods, a higher margin on the superior product would allow the firm to extract additional consumers' surplus.

The hypothesis that the difference in leaded and unleaded regular margins is due to price discrimination is investigated by Severin Borenstein (1989), who argues that purchasers of leaded gas have lower search costs (less brand/location loyalty) than purchasers of unleaded gas. The difference in standard deviations reported in Table 1 is consistent with this argument; more search should reduce the variation in price. The difference in search cost implies differing price elasticities, allowing stations with local market power to price discriminate. Borenstein exploits the fact that, in recent years, an increasing number of stations have stopped carrying leaded gas. As the distribution of leaded gas outlets becomes relatively less dense, the cost of search will rise and the difference in margins should decline. Using panel data on average margins and surveys of consumer purchasing, he finds the difference in retail margins was rising or constant until 1986 when it began to decline, and the change in availability (controlling for differences in income) explains from one- to two-thirds of the drop in the margin difference over the 1986–89 period.

Another way in which stations might be able to price discriminate is through willingness to pay for service. My 1989 paper develops a model in which stations with local market power carrying both full- and self-service gasoline are able to price discriminate, while stations carrying only full-service or only self-service gasoline cannot. As a result, the difference in full- and self-service prices will be larger at stations offering both service types, than the difference across stations offering only one service type. This prediction is tested using the eastern Massachusetts cross-sectional data on station prices and characteristics. Controlling for a variety of station characteristics, these data indicate that the differential is at least 10 cents higher at stations with both service types. Since the wholesale price does not vary by station type, it is difficult to argue that a differential this large can be cost-based. The data base includes location information on each station that is used to control for local deviations in input prices, demand, and competitive environment. The differential remains even when the comparison is restricted to stations within a half-mile radius. Because the larger differential comes from higher full-service prices, the presence of both full-service-only and multiservice stations explains some of the difference in the variance in full- and self-service prices evident in Table 1.

My results and those of Borenstein lend support to the argument that price discrimination, although most commonly modeled in a monopoly setting, occurs in multifirm markets in which market power comes from location and brand preference.

II. Contractual Form

Gasoline is sold in a mixed distribution system, that is, a system where there is a variety of contractual forms for the vertical relationship between retailers and refiners. These forms are usually categorized by the industry as: open dealer (OD) contracts in which the station is owned and operated by an independent dealer who contracts with a refiner for her gasoline supply; lessee dealer (LD) contracts in which the capital at the station is owned by the refiner and leased by the self-employed dealer; and company outlets (CO) at which the capital is owned by the refiner and the station operator is employed by the refiner. Note that the only form in which the refiner can directly control the retail price is the CO. The franchise agreement and capital ownership at LD stations provide greater scope for control of nonprice attributes than does the supply contract signed with ODs.

Ownership of capital allows the supplier to build-in auxiliary services (say, convenience stores, automotive service bays) and determine the capacity of the station (the number of pumps and islands). Ownership of capital also gives the supplier a convenient mechanism for imposing a franchise fee. The franchise contract gives the supplier some additional rights to define the quality of service. Although the allowable scope of these contracts varies from state to state, suppliers can sometimes set the hours of operation, specify maintenance and cleanliness standards, enforce credit card policies, and take some sorts of disciplinary action in response to violations of the contract.

Both state and federal laws and regulations restrict the contracts refiners can offer. These policies have been initiated and supported by organizations representing dealers with LD or, to a lesser extent, OD contracts. The context for this rise in dealer activism is a sharp decline in the number of stations; from 1973 to 1982 the number of stations in the United States declined by over one-third, and the remaining stations have higher sales volumes (U.S. Department of Energy, *Petroleum Market Shares*).

While many analysts view this shift toward fewer and higher volume stations as a rationalization of the distribution system, dealers are understandably unhappy. One result of their efforts is the federal Petroleum Marketing Practices Act (1978) disallowing unilateral dealer terminations (even at contract renewal) except in specifically defined circumstances, and limiting the refiner's ability to extract downstream rents with station-specific franchise fees. To the extent that these restrictions bind, a clear effect would be to push refiners toward company outlets as the preferred contractual form. Perhaps in recognition of this effect, dealer organizations have convinced several states to prohibit company outlets.

My 1990 paper uses the cross-sectional, eastern Massachusetts data to test for relationships between station price and quality characteristics and the choice of contractual form. The refiner's problem is treated as a two-stage decision. First, she chooses the station price and quality characteristics that maximize profit at any given location; she then chooses the contractual form that maximizes profit conditional on the desired station type. Because local market power will lead to double marginalization[3] when some of the rent must be extracted via the wholesale price, stations intended to be high-volume, low-price stations can be run optimally as company outlets.

Stations intended to provide complex services that are difficult to monitor, but affect the brand's reputation, will be run optimally under an LD contract that gives some control over quality to the refiner, but makes the dealer a residual claimant for service income. Using a multinomial choice model, I found (1990) that a low price increases the probability that the station will be a CO relative to either other type, but has no effect on the probability ratio of LD to OD. Stations that provide automotive service, on the other hand, are less likely to be COs and more likely to be LD outlets relative to OD outlets.

Another line of research exploits the regime change when legislation prohibiting COs goes into effect. John Barron and John Umbeck (1984), for example, use data on prices at outlets that were company stores prior to the prohibiting legislation in Maryland, and became LD or OD outlets after the prohibition went into effect. They compare these prices to those at nearby stations over approximately the same time periods. They find that prices at COs were lower relative to nearby stations before prohibition than they were after prohibition. They interpret this result as evidence of double marginalization at stations where refiners cannot control the retail price.

---

[3] G. F. Mathewson and R. A. Winter (1984) argue that in free-entry markets, excessive entry will lead to dealers choosing prices that are too low from the manufacturer's perspective, not prices that are too high. In gasoline retailing, however, refiners can directly control entry.

III. Directions for Further Research

The sparse empirical work on vertical organization suggests that refiners are choosing vertical contracts that minimize total organizational costs (monitoring costs plus the cost of retailers failing to internalize the refiner's profit). It also suggests that one effect of prohibiting COs might be to increase the average retail price. It is less clear what other effects these prohibitions and other restrictions have on the structure and behavior of the retail sector. It may be possible to identify these effects by exploiting the variation in the nature of the restrictions across states.

The studies of retail pricing rely on product differentiation as a source of market power, but they do not directly investigate the nature of spatial and brand competition. The effect on pricing of station density relative to demand could be investigated. Models of differentiation suggest that suppliers of a homogeneous product will want to soften price competition by spatially differentiating, but also want to locate at high-demand areas. It might be that these opposing forces lead stations to cluster in high-demand areas, but differentiate with respect to brand and type of service. Because gas stations (and other retail outlets) have clearly defined locations they appear to be particularly well-suited for investigating these issues.

REFERENCES

Barron, John M. and Umbeck, John R., "The Effects of Different Contractual Arrangements: The Case of Retail Gasoline Markets," *Journal of Law and Economics*, October 1984, *27*, 313–28.

Borenstein, Severin, "Selling Costs and Switching Costs: Explaining Retail Gasoline Margins," Working Paper no. 305, IPPS, University of Michigan, September 1989.

Mathewson, G. F. and Winter, R. A., "An Economic Theory of Vertical Restraints," *Rand Journal of Economics*, Spring, 1984, 28–38.

Shepard, Andrea, "Price Discrimination and Retail Configuration," Working Paper no. 526, MIT Economics Department, June 1989.

______, "Choice of Contractual Form in Retail Markets," MIT Economics Department, 1990.

Slade, Margaret E., "Conjectures, Firm Characteristics, and Market Structure: An Empirical Assessment," *International Journal of Industrial Organization*, December 1986, *4*, 347–69.

______, "Interfirm Rivalry in a Repeated Game: An Empirical Test of Tacit Collusion," *Journal of Industrial Economics*, June 1987, *34*, 499–516.

U.S. Department of Energy, *Petroleum Marketing Monthly*, DOE/EIA-0380, Washington, September 1987.

______, *Petroleum Market Shares*, Series DOE/EIA-066, Washington, 1973–81.

Money-Back Warranties vs. Replacement Warranties: A Simple Comparison

By DUNCAN P. MANN AND JENNIFER P. WISSINK*

Product warranties are one of the most common, time-honored, and visible contractual relationships entered into by consumers. Moreover, under the Uniform Commercial Code, consumers are covered by an implied "warranty of merchantability," even without a written warranty. In this paper, we compare the outcomes of two contrasting types of consumer product warranties: money-back warranties and replacement warranties. Both these types of warranties are ubiquitous. In fact, A. Best and A. R. Andreasen (1977) find that returning an item is the most common action taken by dissatisfied consumers. C. L. Kendall and F. A. Russ (1975), in a study of the consumer packaged goods industry, find that money-back and new merchandise are the two most common responses made by firms to consumer complaints. Yet as our analysis shows, there are significant differences in the incentives they provide firms and in their welfare properties. In particular, we find that the comparison depends crucially on the observability of quality and on the cost of servicing warranties.

Although there are certainly more complicated explanations for warranties, we rely on a simple model of imperfect consumer information in which to embed our comparison of the relative performance of these two forms of warranty. Consumers are confronted by seller moral hazard, that is, either they are unable to discern the quality of a product before purchase or, alternatively, this information is extremely costly for consumers to obtain or for firms to publicly demonstrate. We abstract from issues such as differing degrees of risk aversion, consumer moral hazard, and screening, each of which is frequently the rationale behind warranty provision.[1] Our model is perhaps closest to the model presented in Michael Spence's 1977 article in which he analyzes the provision of warranties in an environment with biased consumer information.

I. Firms, Consumers, and Warranties

To compare these two consumer protection alternatives, we focus on the transactions between a buyer and a seller for a product that, once purchased, either works or fails to work. Throughout the analysis, we take the seller to be a firm and the buyer to be a consumer; however, both parties can be viewed as firms should the reader prefer that interpretation.

The probability that the product works is denoted π, thus $(1-\pi)$ is the probability of failure. The seller of the product influences π through an input, q, that can be thought of as the product's manufactured quality. Firms are all endowed with identical technologies for producing products. Quality has a positive, but diminishing marginal impact on reliability: $\pi_q > 0$, $\pi_{qq} < 0$, and $\pi \leq 1$.[2] The firm's unit cost function for quality, $c(q)$, is strictly convex and does not entail any fixed costs (i.e., $c(0)=0$, $c_q > 0$, and $c_{qq} > 0$), while the technology for producing additional units is constant returns to scale. The functions $c(q)$ and $\pi(q)$ are assumed to be common knowledge, however, their *ex post* values are not publicly verifiable. Each firm chooses quality to maximize profits, that is, firms are assumed to be risk neutral.

---

*Williams College, Williamstown, MA 01267, and Cornell University, Ithaca, NY 14853, respectively.

[1] For example, see Geoffrey Heal (1977) on risk sharing; George Priest (1981), Russell Cooper and Thomas Ross (1985), Nancy Lutz (1989), and our 1988 and forthcoming papers, on consumer moral hazard; and Steven Matthews and John Moore (1987) on screening.

[2] Throughout the paper subscripts refer to derivatives or, where appropriate, partial derivatives.

Consumers are assumed to place a monetary value, V, on the services derived from a working product and attach zero benefit to products that fail. These benefits are identical for all consumers. Consumers are also assumed to be risk neutral with respect to the monetary value of the services derived from the product.

There exist several types of warranties that firms could offer consumers in this simple environment. In this paper, we investigate two in particular, money-back and replacement warranties. A promise by the firm to replace the product if it is found to be defective we term a replacement warranty, while a promise to return all, part, or a multiple of the consumer's purchase price if a product fails is termed a money-back warranty. With a money-back warranty, we assume that the payment of a refund by the firm terminates the firm-consumer relationship. In contrast, a replaced product is assumed to carry the same warranty as the original and may itself be replaced if it fails to work. The consumer is "tied" to the firm until she receives a working product. These differences in warranties lead to different incentives for firms to provide quality and, consequently, affect the consumer's expected net benefits from contracting with the seller.

The processing of warranties is likely to impose some costs on the firms involved. We assume that the payment of a refund for a money-back warranty, or the replacement of a broken product imposes a fixed cost on the firm. There may also be some costs that consumers bear, however; these costs only figure in determining whether or not the consumer exercises the warranty provisions and are ignored in the analysis. The presence of these consumer costs would not alter our results as long as the costs to consumers of exercising a money-back warranty do not differ from the costs incurred with obtaining a replacement.[3]

II. Money-Back Warranties

A money-back warranty is represented by the pair (p, w), where p is the price a consumer pays for a single unit of the product, and w is the refund the consumer receives from the firm if the product fails. After the seller has chosen the product's quality, the product is observed to work or fail. If it fails the warranty is exercised. We assume that a third party can costlessly determine whether or not a product works and enforce the terms of the contract without dispute.[4] However, the input level provided by the firm is not publicly verifiable, and therefore cannot be part of any enforceable contract.

As a first step towards describing an equilibrium contract, consider how the seller's input choice is affected by the terms of the contract. A firm's expected profit from a contract specifying the terms (p, w) is $S(q|p, w) = p - c(q) - [(w+T)(1-\pi(q))/(1+r)]$, where T is the cost of servicing the warranty, and the interest rate is denoted by r. Given the assumptions made on $c(q)$ and $\pi(q)$, along with minor boundary conditions, it is easy to show that a given warranty level will uniquely determine the seller's optimal quality.[5] That is, the firm's optimal quality as a function of the warranty, $\hat{q}(w|r, T)$, is defined implicitly by the relationship $c_q = (w+T)\pi_q/(1+r)$. It is straightforward to show that quality increases with both the warranty level and with the cost of servicing the warranty. Product failures are more costly the higher w and T. Consequently, firms choose to reduce the probability of failure by manufacturing higher quality products.

We assume that the product is sold in a market where a large number of identical firms offer competing contracts. Free entry drives active firms' expected profits to zero. In equilibrium, competition forces producers to supply contracts that maximize the buyer's induced net benefit, subject to a zero expected profit constraint. That is,

---

[3] See Edward Golding (1982) for a model that examines warranties when some consumers do not seek redress.

[4] Several problems develop if product failure is costly to verify. See, for example, Thomas Palfrey and Thomas Romer (1983) and our 1988 paper.

[5] In the interest of brevity, here, and throughout the paper, formal proofs of statements are omitted. They can be obtained from the authors upon request.

$$(1)\quad \max_{w} B = \pi(\hat{q}(w))V - p + (1-\pi(\hat{q}(w)))\,w/(1+r)$$

subject to $p = c(\hat{q}(w)) + (1-\pi(\hat{q}(w)))(w+T)/(1+r)$.

This problem is concave provided $\pi_{qqq} \leq 0$ and $c_{qqq} \geq 0$. The first-order condition for (1), along with $\hat{q}(w|r,T)$, imply that $\hat{w} = V(1+r)$. Knowing $\hat{w}$ determines equilibrium values for $\hat{q}$, $\hat{p}$, and $\hat{B}$. There are several points to note about this contract. Interestingly, it can be shown that in these contracts $\hat{w} > \hat{p}$. Contracts must promise *more* than "satisfaction or your money back" if they are to police sellers' incentives. Also, the seller is forced to bear all the risk as a consequence of his unobservable action. The present discounted value of the buyer's income is the same whether the product works or fails. Finally, the market contract protects the seller against buyer sabotage or dispute. Since $\hat{w} = V(1+r)$, the buyer has no incentive to intentionally break the product or falsely claim that the product failed in order to get the warranty. Therefore, to enforce these contracts an outside party need only ensure that all buyer claims be honored.

For the sake of comparison, consider a situation in which quality is observable while all the contracting surplus continues to go to the buyer. In this case, q will maximize $B = \pi(q)V - c(q)$. Standard assumptions make this maximization problem concave. Let q^* and B^* be the solutions to this exercise. It is straightforward to show that at $T = 0$, $q^* = \hat{q}$, and consequently, $B^* = \hat{B}$. When there is no cost to servicing the warranty, the market warranty completely ameliorates the seller's moral hazard. This result is quite similar to Spence's results for risk-neutral buyers and sellers. In his case, however, the government must set the warranty as a matter of policy since there are consumer misperceptions regarding the function $\pi(q)$.

As T increases it can be shown that $\hat{w}$ remains fixed while $\hat{q}$ increases and $\hat{B}$ decreases. Both q^* and B^* are invariant with respect to T. Not surprisingly, increasing the cost of servicing the warranty reduces the consumer's expected benefits from contracting with the seller and leads to an overproduction of quality relative to the standard of observable quality.

III. Replacement Warranties

A replacement warranty is represented by a price and a promise to furnish the buyer, should her unit fail, with a new unit. Recall that a replacement product is assumed to carry the same warranty as the original and may itself be replaced if it, too, fails to perform.

Consider first the outcome from a replacement warranty assuming observable quality and competitive markets. In this case, quality would be chosen to maximize buyers' expected benefit subject to a zero profit condition for the seller. This yields, after substitution of the zero profit condition, the following problem:

$$(2)\quad \max_{q} B = [\pi(q)V - c(q)] + \sum_{i=1}^{\infty} \frac{[1-\pi(q)]^{i}}{(1+r)^{i}}[\pi(q)V - c(q) - T],$$

which reduces to

$$(3)\quad \max_{q} B = \frac{[\pi(q)V - c(q)](1+r)}{r+\pi(q)} - \frac{[1-\pi(q)]T}{r+\pi(q)}.$$

We assume conditions sufficient for this function to be concave. The first-order condition for the optimal q, denoted q', is

$$(4)\quad c_q = (\pi V - c)\frac{-\pi_q}{(r+\pi)} + \pi_q V + \frac{\pi_q T}{(1+r)} + \frac{(1-\pi)T\pi_q}{(r+\pi)(1+r)}.$$

The implied q' generates an expected benefit, B', for the typical buyer.

If quality is unobservable, the seller's unit expected profit function from a replacement warranty is

$$(5)\quad S = p - c(q) - \sum_{i=1}^{\infty} \frac{[1-\pi(q)]^{i}}{(1+r)^{i}}[c(q)+T].$$

Simplifying the above yields $S = p - [c(q)(1+r)]/[r+\pi(q)] - [T(1-\pi(q))]/[r+\pi(q)]$, which the seller maximizes with respect to q. Once again, conditions sufficient for this function to be concave are assumed. The first-order condition for the optimal q, denoted $\bar{q}$, is

$$(6)\quad c_q = \frac{c\pi_q}{(r+\pi)} + \frac{T\pi_q}{(1+r)} + \frac{(1-\pi)T\pi_q}{(1+r)(r+\pi)}.$$

Given $\bar{q}$, competition will determine the zero profit price, $\bar{p}$, that in turn determines the buyer's expected benefit from the replacement contract, $\bar{B}$.

A number of observations can be made at this point. First, when quality is unobservable and $T = 0$, the market solution with replacement warranties degenerates; the seller's expected profit is maximized at $q = 0$. Intuitively this makes sense. If the firm's only costs are variable production costs, once the seller has captured the buyer's money, then a promise to indefinitely replace junk with junk maximizes profits. Consequently, the price falls to zero and products with replacement warranties are not offered. Assuming $T > 0$ and comparing conditions (4) and (6), it is straightforward to show that $\bar{q} < q'$, implying $\bar{B} < B'$. Once again, lack of observability hurts the buyer. However, in contrast to money-back warranties, when quality is unobservable and firms offer replacement warranties, quality is underprovided relative to the full-information ideal. Whether observable or not, quality will increase with the cost of servicing the replacement warranty. Furthermore, B' decreases with T while $\bar{B}$ increases at first and then decreases.

IV. Money-Back vs. Replacement Warranties

We are now in a position to compare the desirability of money-back vs. replacement warranties. The comparison depends crucially on the observability of quality and the cost of servicing warranties. Consider first how the two alternative contracts compare when quality is observable. Recall that if the seller and buyer can contract directly over quality there is no need for the money-back promise. The buyer and seller simply engage in a one-shot transaction for a specified level of quality. The comparison here is between a one-shot "gamble" or a repeated one. Comparing B^* and B', we find that, in general, there will exist some T below which the replacement contract dominates and above which consumers fair better with the one-shot deal. For small values of T, it can also be shown that $q' < q^*$. When servicing costs are sufficiently small, it may be that buyers are better off with less reliable products that are replaced if defective.

When quality is unobservable, the cross-contract comparison is a bit more complicated. When $T = 0$, it is obvious that the money-back warranty is better. It is uncontested. In general, as T increases from zero, there is a range where the money-back warranty continues to dominate the replacement contract. For an intermediate range of values for T, the replacement contract dominates the money-back contract, only to eventually be taken over once again by the money-back contract for relatively large values of T.

This pattern is intuitive and consistent with the discussions in the previous sections. When the costs of servicing the warranty are quite small, replacement warranties serve as a relatively weak shirking deterrent. Buyers can expect poorer quality products from establishments that can easily replace broken merchandise. At the other end of the spectrum, when the costs of servicing the warranty are relatively high, replacement warranties are inferior because the expected cost of repeatedly handling claims is passed back to the buyer through the price, thereby diminishing the buyer's expected benefits from the replacement contract to a level below the level expected when offered a money-back guarantee. How this pattern is affected by changes in our simplifying assumptions is part of our present research program.

REFERENCES

Best A. and Andreasen, A. R., "Consumer Response to Unsatisfactory Purchases: A Survey of Perceiving Defects, Voicing Complaints and Obtaining Redress," *Journal of Law and Society*, Spring 1977, *4*, 701–42.

Cooper, Russell and Ross, Thomas W., "Product Warranties and Double Moral Hazard," *Rand Journal of Economics*, Spring 1985, *16*, 103–13.

Golding, Edward L., "Warranties as Signals of Product Quality When Some Consumers do not Seek Redress," FTC Working Paper no. 69, Washington, 1982.

Heal, Geoffrey M., "Guarantees and Risk Sharing," *Review of Economic Studies*, October 1977, *44*, 549–60.

Kendall, C. L. and Russ, F. A., "Warranty and Complaint Policies: An Opportunity for Marketing Management," *Journal of Marketing*, April 1975, *42*, 36–43.

Lutz, Nancy A., "Warranties as Signals under Consumer Moral Hazard," *Rand Journal of Economics*, Summer 1989, *20*, 239–55.

Mann, Duncan, P. and Wissink, Jennifer P., "Money Back Contracts with Double Moral Hazard," *Rand Journal of Economics*, Summer 1988, *19*, 285–92.

______ **and** ______ , "Hidden Actions and Hidden Characteristics in Warranty Markets," *International Journal of Industrial Organization*, forthcoming.

Matthews, Steven and Moore, John, "Monopoly Provision of Quality and Warranties: An Exploration in the Theory of Multidimensional Screening," *Econometrica*, March 1987, *55*, 441–67.

Palfrey, Thomas R. and Romer, Thomas, "Warranties, Performance and Resolution of Buyer-Seller Disputes," *Bell Journal of Economics*, Spring 1983, *14*, 97–114.

Priest, George L., "A Theory of Consumer Product Warranty," *Yale Law Journal*, May 1981, *90*, 1297–352.

Spence, Michael, "Consumer Misperceptions, Product Failure and Product Liability," *Review of Economic Studies*, October 1977, *44*, 561–72.

Educating and Training New Economics Ph.D.s: How Good a Job Are We Doing?

By W. LEE HANSEN*

How good a job is the economics profession doing educating and training new Ph.D.s, and thereby shaping the nature and direction of current and future research in economics? Have our graduate training programs drifted away from their central task of producing new Ph.D.s who can perform a wide variety of teaching and research activities in both academic institutions and elsewhere? Has economics training become an increasingly theoretical, technique-driven enterprise, giving short shrift to examining real-world economic problems with important empirical, institutional, and policy dimensions?

These and similar questions led the AEA's Executive Committee in early 1988 to establish the Commission on Graduate Education in Economics. The Commission's purpose was to examine the structure and content of graduate education in economics, and to make whatever recommendations seemed appropriate. Its recommendations are to be directed to the economics profession and particularly to economics departments with Ph.D. programs. Because the Commission is an independent body, its recommendations will be the responsibility of its members; neither its findings nor its recommendations will be endorsed by the American Economic Association. The Commission hopes to complete its work within the next several months. Its findings and recommendations will appear later this year, with a longer report on the Commission's work to follow.

Over the past 15 months, the Commission has engaged in a variety of projects. Extensive information obtained from economics departments is being analyzed, as are the results of a comprehensive set of surveys. In addition, various special studies are in progress.

The preliminary results presented here are based on surveys of faculty members, graduate students in their second year and beyond, and recent Ph.D.s, all of whom were selected from a sample of departments with graduate programs that were among those ranked in the 1982 study by the National Academy of Sciences. Questionnaires were mailed in March and April, and yielded response rates ranging from 60 to 70 percent. The respondents appear to be reasonably representative of the underlying population groups.

The focus of this report is on how and how well graduate students are educated and trained. Through a mixture of questions that tried to capture what is and what should be, it becomes possible to pinpoint the concerns of the active participants (faculty members

---

*University of Wisconsin, Madison, WI 53706, and Executive Secretary of the AEA Commission on Graduate Education in Economics, chaired by Anne O. Krueger. Three other members of the Commission (Alan S. Blinder, Claudia Goldin, and T. Paul Schultz) comment on other aspects of the Commission's work, and Robert Solow (not a member of the Commission) comments specifically on this paper. Other members of the Commission are Kenneth Arrow, Olivier Jean Blanchard, Edward Leamer, Robert Lucas, John Panzar, Rudolph Penner, Joseph Stiglitz, and Lawrence Summers.

The Commission is indebted to the Andrew W. Mellon Foundation, the Alfred P. Sloan Foundation, and the National Science Foundation for funding its work. It also expresses its appreciation to the many departments, faculty members, graduate students, recent Ph.D.s, undergraduate chairs, employers, and undergraduate majors who responded to its surveys. The Executive Secretary offers special acknowledgment to Edward Bird and Thomas Buchmueller for research assistance, to Jeffery Dominitz and Nancy Omernik for assistance on several special projects, and to Suzanne Vinmans for secretarial support.

and graduate students) as well as the concerns of former graduate students, some of whom are themselves now faculty members in these same graduate departments.

In some instances, results are provided for different groups of graduate programs based on departmental quality. Thus, tier 1 consists of the top 6 economics Ph.D. programs; tier 2 includes the next 9 programs, tier 3 the next 15, tier 4 the next 20, and tier 5 the remaining 42 ranked programs. The results for recent Ph.D.s focus on the 1977–78 doctoral cohort to illustrate how increased work force experience influences views about graduate education and whether it should be changed.

I. Is There Need for Change?

There is considerable sentiment both among faculty members and graduate students, and to a lesser extent among recent Ph.D.s, that changes are needed in the content and structure of Ph.D. programs in economics. Among faculty members, 60 percent expressed support for changing the Ph.D. programs in their own departments, as compared to 30 percent who opposed change; another 10 percent did not know or were not sure. When asked about the need for changes in Ph.D. programs generally, support dropped off to 50 percent, with another 25 percent opposing change, and the remaining 25 percent not responding or indicating that they did not know. Once the "don't know" group was excluded, roughly two-thirds of those faculty members who expressed themselves one way or the other favored changes in their own or other programs.

The proportions of graduate students favoring change was slightly greater, with about two-thirds wanting change in the programs in which they enrolled. Of those who responded either yes or no, almost 80 percent favored change. Support for change among recent Ph.D.s was weaker. When asked about the need for changes in the programs from which they obtained their Ph.D.s, approximately 45 percent favored change; a similar result obtained for their current Ph.D. department. If those recent Ph.D.s who responded "don't know" are excluded, then only a slight majority of recent Ph.D. recipients favored change.

The responses by department quality tier are illuminating. Faculty members at top-tier departments were least interested in change, whereas those at bottom-tier departments were most interested in change. Among students, only those at bottom-tier departments exhibited any noticeably greater support for change.

II. Acquiring Knowledge and Skills

A central objective of the surveys was to find out more about what graduate students learn, and what they should learn. In general, graduate students believe that, in the first two years of study, they acquire the essential knowledge of economics they need. They do not, however, believe they acquire the complementary skills that are needed by economists.

Knowledge. To find out what graduate students are learning, all three groups of respondents were asked to rank the relative importance now given, and that should be given, to each of six different types of knowledge. The list contained economic theory, econometrics, institutions and history, literature, applications, and empirical economics. The results indicate substantial agreement across the three groups, and a general satisfaction with the current emphasis (i.e., on what is and what should be). For faculty members and for graduate students, the resulting order of the rankings on both what is and what should be comes out about the way most economists would have expected. Economic theory leads the way, followed in order by econometrics, empirical economics, and applications; institutions and literature are at the bottom of the list.

The rankings of graduate students should not be surprising; they quite naturally mimic the emphasis conveyed by the faculty. The rankings provided by recent Ph.D.s were also the same. This applied not only to their ranking of departments from which they obtained their degrees, but also to graduate departments in which they were now situated.

This evidence suggests that faculty members, graduate students, and recent Ph.D.s see nothing seriously wrong with the relative emphasis given to the six different types of knowledge.

Skills. Graduate education imparts both knowledge and skills. Based on a list of skills developed for the study, respondents were asked to rank the importance that is given, and should be given, to seven different skills. The list contained analytics, mathematics, critical judgement, applications, computation, creativity, and communications. In describing what is currently important in graduate programs, faculty members and graduate students once again gave similar rankings to these skills; so also did recent Ph.D.s for the programs from which they obtained their doctorates. In fact, the rankings for all three groups follow the order of the list of skills just given.

The results differed considerably when respondents were asked to indicate the importance that should be assigned to these skills. While analytics remained at the top of the list, mathematics and computation declined sharply in importance, moving to the bottom of the list for faculty members and close to the bottom for graduate students. Meanwhile, creativity advanced to second position for faculty members, and to third position for graduate students. The skill of application remained at the middle of both rankings.

Recent Ph.D.s, in evaluating which skills had been important in their own Ph.D. training, gave similar responses. But, rather than posing the "what should be" question to them, it seemed more appropriate to ask them to rank the importance of these skills in their current jobs (that ranged from employment in graduate economics programs to positions in government and business). Again, the importance of mathematics dropped off, falling from second to seventh place in importance on the job, even for those employed in departments with Ph.D. programs. Meanwhile, the skills of communication and creativity moved up. These changes indicate quite a different focus between the training of recent Ph.D.s and the work they subsequently do.

These results also show that the mix of skills used on the job corresponds more closely to expressions by faculty members and graduate students about the mix of skills that should be important in graduate education, than about the mix of skills currently emphasized in that training. Indeed, both faculty members and graduate students can be interpreted as sensing that the mix of needed skills differs from the skills being developed in graduate training.

To probe this matter, faculty members were asked about the importance that is, and should be, given to these same skills in Ph.D. dissertations. Creativity and critical judgement scored the largest gains in what should be emphasized (relative to those skills that are currently emphasized in graduate training), while mathematics and computation registered the largest declines. In other words, the mix of skills needed for both dissertation research and subsequent employment differs from the mix of skills currently emphasized in graduate programs.

As further evidence of the mismatch that prevails, faculty members were asked to evaluate, on a scale of excellent to poor, the quality of the skills possessed by recent newly hired Ph.D.s in their departments. Eighty-five percent of faculty respondents gave new hires excellent or good evaluations for analytical ability and mathematics. Slightly less than two-thirds gave similar evaluations of their skills in computation and critical judgement. On the down side, less than one-half the newly hired Ph.D.s scored that well on communication, and about one-half scored that well on creativity and on application. Not surprisingly, newly hired Ph.D.s exhibit the very same skills that are being developed among currently enrolled graduate students in these institutions. These are not, however, the skills that economists should have, as viewed either by faculty members and graduate students, or by recent Ph.D.s who practice these skills in the marketplace. Moreover, the skills of sizeable numbers of new Ph.D.s are judged by faculty members to be not very highly developed.

Summary. Two conclusions emerge thus far. First, there appears to be general agreement about what knowledge is and what knowledge should be emphasized in graduate education. Second, there is considerable divergence between the importance that is and should be given in graduate training to the various skills required for economists. Both faculty members and graduate students think the emphasis should differ, by focusing on skills similar to those on the job by recent graduates. Yet, faculty members and Ph.D.-granting departments continue to hire new Ph.D.s with the same skills their own Ph.D. programs emphasize, rather than new Ph.D.s with the skills that are used in the labor market. Of course, it could be that departments are forced to hire the more narrowly skilled new Ph.D.s because no others exist.

III. Developing Proficiencies

Another approach is to focus on what might, for lack of a better term, be called proficiencies. These proficiencies try to capture various combinations of knowledge and skills that will permit graduate students to undertake a broader set of tasks essential to their development as research economists. Expressed another way,.development of these proficiencies can be viewed as an intermediate goal of Ph.D. training, with the ultimate goal being that of producing accomplished Ph.D. economists.

A list of these proficiencies was developed to make this assessment. The list ranges from relatively narrow activities such as readying students for the next steps in the formal parts of their Ph.D. programs (i.e., preparing them for field courses and preparing them for preliminary examinations) to broader activities that serve them in their dissertation research and beyond (i.e., providing rigorous training in economic theory, deriving implications of mathematical models, using theory in empirical applications, applying theory to real world problems), and to the still more encompassing activity of research (preparing students to carry out independent research).

The Graduate Student View. Graduate students indicated substantial variability in rating the contribution of their courses to the development of these proficiencies. The percentages who indicated that all or many of their core theory courses placed substantial emphasis on preparing students for preliminary examinations and field courses were 68 and 49 percent, respectively. These courses did much better on the theory and on the mathematics front, with 94 and 60 percent, respectively, indicating that all or many of their core courses contributed to these objectives. Core courses were said to give less emphasis on using-theory and on applying-theory; only 15 and 14 percent, respectively, indicated that all or most of their core courses placed substantial emphasis on these activities. These courses do only slightly better in the emphasis they placed on independent research, but still only 20 percent of the students indicated that substantial emphasis was given to this objective in all or most of their core courses.

As compared to core courses, fewer students said that field courses placed substantial emphasis on theory and on mathematics (62 and 63 percent), and more respondents said that these courses emphasized both using-theory and applying-theory (45 and 50 percent). In addition, more indicated these courses prepared students to do independent research (50 percent). While the increased emphasis given to the last three items (using, applying, and preparing) is commendable, it is apparent that field courses do not provide strong linkages between courses and dissertation writing. Indeed, approximately one-half the graduate student respondents indicated that few or none of their field courses gave substantial emphasis to preparing them to use theory in empirical applications, to apply theory to real-world problems, or to do independent research. Thus, work remains to be done in bolstering the role of field courses, and in forging closer links with the important first-year core courses, thereby laying a stronger foundation for dissertation research.

For the most part, the impact of the core theory courses on students in different tiers of institutions is surprisingly similar. The same is true of the field courses, except that lower-tier institutions appear to put more emphasis on applying theory to real-world problems. Perhaps that is what those programs emphasize.

The above results are somewhat perplexing in view of how students rate the quality of their instruction. Generally, graduate students are pleased with the instruction they receive. Approximately three-quarters of them describe the quality of instruction in the field courses as excellent or good. In the core theory and econometrics courses, the proportions are only slightly lower, about two-thirds. As a general rule, the quality of instruction drops off a bit at the lower-tier institutions. The other side of this story is that at least in the core theory courses, one-third of the graduate students describe the instruction they receive as either fair or poor. This proportion seems large enough to merit some concern, particularly in lower-tier institutions where the frequency of poor evaluations is greater.

The burden of the evidence is that, while courses are effective in achieving concrete specific objectives (such as preparing graduate students for the next steps in their formal Ph.D. programs), they are weak in helping students acquire the proficiencies they will need to both write their dissertations and to succeed as practicing economists. This is not surprising; the former is much easier to achieve than the latter. Despite this weakness, graduate students seem to be generous, perhaps overly generous, in assessing the quality of the instruction they receive.

The Faculty View. Faculty members were asked to evaluate not what the various courses did or tried to do, but rather the proficiencies achieved by students. Faculty members were harsher in their assessment of a more restricted set of proficiencies. They indicated that, by the time of their comprehensive examinations, many graduate students had not gained proficiency in using the knowledge and skills they had been exposed to, or developed in their course work. At the high extreme, 58 percent of faculty members indicated that all or most students (as compared with some, few, or none) were well-grounded in economic theory. This rather low figure contrasts with the favorable graduate student view of their core courses, and it comes as a surprise, because of the strong emphasis given to theory in the core graduate courses; on the other hand, faculty respondents may have interpreted this as implying a level of understanding that only theory majors could be expected to demonstrate. The extent to which faculty members think graduate students are good at deriving the economic implications of mathematical models is also lower than might have been surmised from the graduate student responses.

In sharp contrast, only 14 percent of faculty members could say that, by the time students complete their comprehensive examinations, most or all of them were good at using theory in empirical applications or at applying theory to the real world. Only 13 percent of faculty members indicated that most or all students at this stage were good at doing independent research.

One concludes from the combination of graduate student and faculty responses that the first two years of graduate work focus on helping students to learn to do what is done in the formal structured parts of their programs. This stands in contrast to sharpening the ability of students to undertake independent research and honing their proficiencies in using the theory in empirical applications, and in applying theory to actual problems. Even then, a considerable gap remains between what students think their course work does for them on the mathematics and theory side, and the assessment provided by faculty members.

IV. Preparing for Independent Dissertation Research

Graduate training, according to both faculty members and graduate students, does too little to prepare students for undertaking independent research, as already indicated.

When graduate students were queried about the effect of various activities in preparing them for dissertation research, the results were somewhat more favorable than those just presented. Approximately one-half the graduate students reported their core theory courses prepared them well or very well; almost three-quarters of the respondents gave this response for field courses; and about 60 percent did so for their econometrics courses. What may be surprising is that so many students (about one-half) thought that studying for preliminary examinations helped prepare them for writing their dissertations.

The experience obtained as a research assistant might have been expected to help prepare students for dissertation research. However, less than one-half the respondents who had research assistantships reported this experience had been useful in that respect. Interestingly, about one-third of the respondents who had been teaching assistants said that this experience had been useful in their dissertation work. One wonders why research assistantships do not contribute more to developing research skills. Might it be that research assistants are given highly specific tasks that do not sharpen their skills at doing independent research? Or, might it be that their interests and those of their faculty supervisors are not well matched?

Generally similar questions were asked of recent Ph.D.s. This group thought many of these same activities had been more important in preparing them to write their dissertations. Seventy-five percent rated their core theory courses as preparing them well or very well, followed closely by 81 percent for their field courses. More than one-half of those who worked as research assistants indicated the beneficial effect on subsequent dissertation research; this was almost three times the number of teaching assistants who gave similar responses. In addition, more than one-half the group thought that seminars they had attended were beneficial.

To the extent that a Ph.D. thesis requires substantial writing, it is surprising that graduate students were not required to do more writing either in their courses or in related activities. Graduate students report writing a total of approximately two papers during their first two years, and a total of approximately three papers in their first three years of graduate education; unfortunately, we have no indication of the nature of these papers—whether they were short or long, literature reviews or something more substantial, or theoretical pieces or empirical studies. The limited amount of writing done implies that faculty members do not regard such activity as an integral part of instruction, or that they simply do not want to spend time reading and commenting on student papers. The fact that graduate students who are in the process of being trained to do research write so little may be another reason why they are so poorly prepared to undertake dissertation level research.

Other measures of research involvement are no more encouraging. The total number of seminars, workshops, or other presentations of original research that graduate students reported they attended during the first two years reached about 3.5, and then leveled off at 4 during the third year. Making presentations of their research was even less frequent, with only one-half the students reporting they had made an oral presentation by their third year, and with more than one-quarter of all students reporting they had still not made an oral presentation by their fourth year.

The low level of activity shown by these crude indicators of research involvement are not inconsistent with the evidence already presented. Neither courses nor other aspects of graduate training offer great stimulus to developing the capacity to do independent research.

To follow up on this point, recent Ph.D.s were asked whether they had, prior to starting their dissertation, completed a research paper on the same or a similar topic. About one-third of the respondents reported having written such a paper; this means that two-thirds of those who completed Ph.D.s launched their dissertation-writing efforts without having had the benefit of what might be called a trial run at research.

The absence of more research activity training and research is difficult to understand. Faculty members reported discussing their own research in their classes; more than one-half said they did this often or very often, particularly at tier-1 institutions. Moreover, 86 percent of faculty members reported having presented their research in some kind of departmental forum during the past year. Of course, the volume of their own research activity was high, with the average faculty member reporting having published four papers in scholarly journals over the previous two years. In short, some kind of research environment obviously exists. Unfortunately, we do not know enough about its quality to understand what accounts for the low volume of graduate student research preparation and activity in light of the substantial amount of faculty research activity.

V. Selecting a Dissertation Topic

Graduate students experience substantial difficulty narrowing and then focusing on a dissertation topic, and later, converting their topic into an acceptable dissertation that can be completed in timely fashion. With the lengthening time to complete the Ph.D. (now about 6.5 years) and the now shorter time to complete the preliminary examinations, the "black hole" of graduate training (the Ph.D. dissertation writing stage) amounts to about 4.5 years. While some of this time no doubt reflects delays because people took jobs before their dissertations were completed (this was true for about one-half of the 1977–78 cohort of new Ph.D.s), the major difficulty seems to lie in identifying the idea for the thesis. Survey responses show that less than 10 percent of the graduate students identified the subject of their dissertation research by the end of the second year; another one-half identified their topic during the third year. Yet, more than 40 percent had not yet passed this critical stage by the beginning of the fourth year; of this group, 16 percent had not found a topic until after the fourth year.

How do students get their ideas? Most students credit "their own research, reading, and/or thinking" as the major source of research ideas for their theses. Faculty members apparently play a much less important role, with only 20 percent of these writers finding their inspiration from a faculty advisor. The activity that occupies most student time, namely, classroom instruction, contributes relatively little, with only 12 percent of the respondents citing class lectures, readings, and the like as a source of ideas or inspiration.

The experience of recent cohorts of economics Ph.D.s differed. Typically, they identified their dissertation topics in their third year, though between 20 and 30 percent reported having settled on a topic in their second year or earlier. In contrast to current graduate students, much larger proportions of recent Ph.D.s gave credit to their faculty advisers. It is interesting to note that the 1987–88 cohort of Ph.D.s indicated, as did current graduate students, that their own ideas, reading, and thinking were the most important source of ideas. The fact that students rely so heavily on their own efforts raises interesting questions. Do they do this because faculty members are reluctant to suggest topics, want to push more of the burden of idea-generation onto students, or are so busy with their own research they have little time for this activity?

Whether and how rapidly students complete their dissertations depends on a variety of factors. Many possible obstacles were posed in the survey questionnaire, such as difficulty of the topic, work as a teaching assistant, inadequate undergraduate preparation, and so on. Many of these obstacles were present, and it appears that between one- and two-thirds of the affected students believed these obstacles delayed their completion. Of particular importance were the standard barriers: work as a teaching assistant, work as a research assistant, financial problems, and family obligations. Other important factors included the difficulty of the thesis topic, problems in obtaining data, changes in dissertation topics, and absence of the major professor. Structural problems also existed; among them, inadequate undergraduate preparation and poor linkage among the components of the graduate program. These obstacles are to be expected in light of the earlier results.

VI. Quality of Recent Ph.D. Dissertations

Doctoral dissertations not only demonstrate the possession of acceptable research skills, but they also embody a contribution (some would argue a significant contribution) to new knowledge. What contributes new knowledge remains unclear—basically, it is what a doctoral committee of between three and five people accepts to fulfill the dissertation requirement for the doctorate.

The quality of the Ph.D. dissertations in economics was not viewed as particularly impressive by faculty dissertation supervisors. This finding emerged from a question asking faculty members to characterize different dimensions of the dissertations they had recently supervised. Based on this list, 68 percent of faculty members agreed that most or all dissertations were a good training instrument. Roughly one-half the faculty respondents thought that all or most dissertations applied economics to real-world problems, or were well-grounded in economic theory. Meanwhile, 40 percent suggested that most or all dissertations offered good empirical evidence to support economic theory, whereas only one-third indicated this applied to showing clearly the implications of mathematical models. Finally, only one-third of the faculty supervisors could say that all or most dissertations were well written. Much of this will strike readers as faint praise.

Most intriguing was the finding that only 21 percent of faculty respondents could agree that most or all of these dissertations were significant contributions to knowledge. Interpreting this is difficult. Though dissertations presumably represent contributions of new knowledge, the inclusion of the adjective "significant" might have produced the 21 percent figure. Alternatively, few dissertations in any field may make significant contributions to new knowledge; hence, the 21 percent figure could be high relative to that in other disciplines. Still, the assessment of dissertations is disquieting.

When faculty members were asked to assign responsibility for the generally low estimate of the quality of Ph.D. dissertations, 23 percent put the blame on students, 13 percent put the blame on the graduate program, and 62 percent thought both sides were at fault. Some minor differences appear among respondents at several tiers. Respondents from the bottom two tiers were more likely to attribute the blame to students, while the opposite was true for respondents at higher tier departments.

VII. Conclusion

These preliminary results (and they represent only a small fraction of the findings) indicate the existence of several important problems in how we educate and train new economics Ph.D.s in our doctoral economics programs. They suggest the need for some important alterations in both the structure and content of graduate education. Finding out how extensive such alterations should be will have to wait until the Commission completes its work and submits its report.

DISCUSSION

ALAN S. BLINDER, Princeton University: Our surveys found widespread dissatisfaction with graduate education in economics—and a fairly consistent pattern of complaints. Both students and faculty find economics obsessed with technique over substance, or too theoretical, or too mathematical, or insufficiently connected with the real world, or too removed from policy and institutional context. The problem seems not to be that we teach poorly; so it must be with what we teach.

What *do* we teach in the core curriculum? We asked graduate students whether all, many, some, or none of their core theory courses "placed substantial emphasis" on each of seven objectives. Unsurprisingly and quite appropriately, these courses devote the most time to "providing rigorous training in economic theory." About 74 percent of the students said all or many of their core courses emphasized that.

The next highest score (68 percent) came on "preparing students for preliminary exams." The impression I carried away from the many macro and micro theory exams the Commission examined is that they tested mathematical puzzle-solving ability, not substantive knowledge about economics. In fact, when Bob Lucas went over the macro exams systematically, he isolated five canonical questions, variants of which appeared over and over again. In my judgement, all tested technique, not substance.

The last objective that received strong emphasis was "deriving economic implications from mathematical models." About 60 percent of the students reported that all or many of their core courses emphasized this, as they should.

The rest of the news was not good. First, only 49 percent of the students said their core courses emphasized "preparing students for field courses." If the core is supposed to equip students with tools to be used in field courses to acquire substantive knowledge about economics, these responses are a bit disconcerting.

Worse yet, only 20 percent of the students said their core courses emphasized "preparing students to do independent research." Since Ph.D. programs are designed to produce researchers, this is worrisome. Of course, one might make a division-of-labor argument that field courses, not the core, should prepare students for research.

However, even this would not excuse the finding that only 14 percent of the students report that their core courses put substantial emphasis on "applying economic theory to real-world problems." This strikes me as a devastating critique. To what, one wonders, is all this theory being applied? The problem, by the way, seems not to be that graduate education is too mathematical. Both faculty and students agree that core theory courses use roughly the right level of mathematics. Rather, the problem is that we train our mathematical artillery on imaginary rather than real targets.

Finally, we found that only about 14 percent of the students thought that all or many of their core courses placed substantial emphasis on "using economic theory in empirical applications." Since these are economic *theory* courses, however, such inattention is neither surprising nor particularly damning.

One last problem with the core curriculum arose. We found that some of the syllabi displayed a disturbingly scatter-shot nature. Students in many universities apparently receive a highly selective education—especially in macroeconomics. Such unevenness in coverage is perhaps to be expected, maybe even encouraged, in field courses. But some commissioners find it disturbing that there is so much idiosyncratic noise in courses that allegedly provide the basic tools.

CLAUDIA GOLDIN, University of Pennsylvania: Fields serve various roles in graduate education. For one, they enable an expanded teaching of theory and econometrics. For another, they allow an exploration of a particular set of facts and issues. They may also provide a place to address many of the deficiencies in the economics curriculum identified by the respondents of the COGEE surveys.

Only 50 percent of the current students surveyed thought that many or all of their field courses applied theory to real-world problems, and just 45 percent thought many or all of their field courses used theory in empirical applications. Only 50 percent thought many or all of their field courses prepared them for independent research. But about 80 percent of recent Ph.D.s thought fields prepared them well for dissertation work—a greater percentage than for any other factor mentioned. Three-quarters of current students believed the quality of instruction in field courses was good to excellent, compared with two-thirds for the core courses. Thus, we get very good marks in the teaching of the fields and they appear to provide the best preparation we offer for a writing dissertation. But, many of the courses do not apply theory to real issues and do not use theory in an empirical manner, and these were among the primary concerns of respondents to a set of open-ended questions on the survey.

What is wrong with the fields emerges with greater clarity from the written comments of the respondents concerning possible changes in the curriculum. The following five statements appeared with considerable frequency. Students claimed: 1) they write too few papers and engage in too little research activity; 2) their training is too narrow and the intellectual content of graduate education is low; 3) they want an applications-oriented approach; 4) there should be more emphasis on policy problems and institutions; and 5) they want courses infused with more economic history and history of economic thought. Faculty responded similarly. They, too, thought students should write more papers, and that there was a need for more emphasis on real-world problems, policy, historical experiences, and the application of theory. And they, too, expressed the sentiment that both the economy and economic theory have meaningful histories that must be taught to students.

Therefore, many of the most serious deficiencies with graduate education in economics may concern the teaching of the field courses. Where else can we impart to students knowledge of real-world issues, facts, institutions, and data? What better place to write papers and begin the creative research process that is the ultimate goal of many?

To assess why problems have arisen in the teaching of the fields, the Commission examined recent field examinations from a wide array of schools. We concluded that some of the fields do appear to teach, or at least reward, creativity, problem solving, knowledge of data, facts, institutions, and policy issues. Labor economics, in particular, seemed successful in achieving these goals, and public finance and development did to some extent. Teaching in fields that have undergone vast changes over the last decade did not seem as effective. Such fields tend to emphasize theory with virtually no empirical, and only tenuous policy, connections. Part of the problem in some of these fields, such as Industrial Organization, appears to be a relative dearth of a recent empirical literature. Some fields seem too large to be treated adequately on one examination. International, for example, covers trade theory and international finance, and open-economy macro has become an expansive subdiscipline. Other fields, such as development, have become fragmented, with some schools emphasizing the micro side and others the trade end. In sum, less than one-half the fields appear to meet the five considerations mentioned above.

We have also examined current curriculum innovations. One is the use of field papers for one or two of the usual field examinations required of students. Such an innovation, currently being tried in at least one tier-2 department, would engage students earlier in the research and writing process. The difficulty, however, is that examinations have obvious scale economies that papers do not. Another innovation is the multifield seminar. Such a research forum would involve several faculty members from diverse fields and would teach students how research is done. One department has initiated such a seminar on the subject of aging, drawing on faculty from the fields of public finance, labor, economic history, and development.

The COGEE study is a truly extraordinary one. Few academic disciplines, if any, have subjected themselves to such rigorous self-examination. Some of the professions have, and they, too, have come to the conclusion that there may be too much theory and too little practice in their training programs.

T. PAUL SCHULTZ, Yale University: My remarks are intended to place the evaluation of graduate education in economics in a broader context of the supply of and demand for economists.

How has the price of economists changed? General wage trends of the last two decades are attributed to macroeconomic conditions and changes in the age/education composition of the U.S. labor force. The level of real earnings of U.S. workers has not increased from the end of the 1960s to the mid-1980s. Increases in output per capita that did occur are traced to increases in participation rates, mostly by women, and payments to other factors than labor. There has also been a decline in the 1970s in the earnings of college-educated workers relative to those who only completed high school.

To describe the earnings of a relatively small group, such as Ph.D. economists, it is necessary to consult specialized surveys. The National Surveys of Natural and Social Scientists and Engineers are thus being analyzed by June O'Neill and Nachum Sicherman for the Commission. These surveys sampled the 1970 and 1980 decennial censuses to obtain current information on income and employment of scientists in 1972, 1974, 1976, 1978, and 1982, 1984, and 1986.

Doctoral economists experienced a decline in real earnings from 1972 to 1978 and an increase from 1982 to 1986, with only a small overall change in this 14-year period. A similar pattern emerges for other scientific disciplines, where only engineering gained about 5 percent relative to economics, and psychology and other social sciences lost 5 percent. Most of the downward drift in real earnings among economists in the 1970s was concentrated in academia and, not surprisingly, the proportion of new Ph.D.s employed in academia declined, from 69 percent in 1958–62 to 56 percent in 1978–87. The share of new Ph.D.s finding their first job in government also declined in the 1980s, as did salaries in government relative to those in academia or business. To abstract from changes in the age composition of the comparison groups, years of experience after receipt of degree and several other characteristics, such as race and sex, are held constant. Individuals, therefore, may have continued to experience growth in real incomes if the payoff to their experience exceeded the declining trend in earnings. The annual salaries of women doctoral economists advanced relative to men. This gain was concentrated in the nonacademic sector where the gap was initially larger. Apparently, the derived demand for Ph.D. economists in the nonacademic economy was sufficient to maintain average salaries of economists at the relatively high levels they achieved during the 1960s, relative to other scientific disciplines.

Earnings of economists with a BA improved compared with those of all college graduates, but not enough to avoid the decline in real wages that marked the 1970s. Further analysis is needed to assess how economists have fared relative to close substitutes, such as business and policy Ph.D.s and MAs.

The number of economics Ph.D.s awarded in the United States was approximately constant during the 1950s, and then tripled during the 1960s, from about 250 per year to 750 per year, where it has remained. Less than 100 Ph.D.s were awarded in businesses in 1951–52, but more than 800 have been awarded annually since 1970. Agricultural economics Ph.D.s also increased rapidly in the 1960s, but declined thereafter to stabilize at around 160 per year since 1975. Public administration or policy schools increased their production of Ph.D.s until they reached 100 in 1975, decreasing slightly thereafter. Thus, allied fields of economics have also experienced growth in Ph.D. programs, but the only comparably sized expansion occurred in business and, to a much lesser extent, in public administration or policy.

In contrast, graduate schools awarded a relatively small number of economics MAs (i.e., 1,937 in 1985–86), which is dwarfed by the 67,137 in business and 4,729 in public administration. Undergraduate majors in economics have also continued to grow modestly since 1965 after doubling during the 1960s. Nonetheless, the number of BAs in economics was still only 21,602, less than one-tenth of those in business in years 1985–86.

All tiers (as defined in Hansen's paper) of economic graduate schools at least doubled their output of Ph.D.s during the demographic boom of the 1960s, but there has been little overall growth in Ph.D.s awarded after 1968–72. However, in this latter period the share of Ph.D.s awarded to U.S. citizens decreased from 73 to 56 percent in 1983–87, and is probably less than one-half in today's entering classes. Only U.S. engineering schools graduate a much larger fraction of foreign students.

Has this change in student body created new opportunities or caused problems for graduate economics education? Are foreign students being admitted because better U.S. students have opted for other prospering fields, such as business, law, or medicine? Or has demand by foreign students for U.S. economics training increased, due to U.S. leadership in this increasingly homogeneous, international, and technical field? Both hypotheses warrant more study. Several indicators suggest that this growth in foreign students is a sign of disciplinary strength, but it probably encourages the tendencies, noted by Hansen, to emphasize mathematical methods and to reduce the requirements for proficiency in written expressions in U.S. graduate education in economics.

Foreign students are about evenly distributed across the various tiers of graduate schools. While the median time from registration in an economics Ph.D. program to receipt of Ph.D. has increased from about 5 to 6 years in the last 15 years, this trend is also evident throughout graduate education in the United States. This delay in the completion of Ph.D. programs is slightly less pronounced among foreign students than U.S. citizens in economics. The GRE scores are a commonly used indicator of student quality. The GRE scores for mathematical aptitude and economics are not noticeably different for foreign than for U.S. citizens, nor are trends different over the last decade or across graduate school tiers. Verbal GRE scores are lower for foreign students, as might be expected, and this may point to a special need in U.S. economics graduate instruction.

Do the foreign Ph.D. students seek work in the United States or return to jobs abroad? The National Research Council's files on new Ph.D.s report information on first jobs after the doctorate. These files show that 2–3 percent of U.S. citizens take their first job abroad, whereas about 10 percent of those with permanent U.S. visas do, and 60 to 70 percent with temporary U.S. visas take first jobs abroad. This last group has grown from 17 percent of U.S. Ph.D.s in 1968–72 to 28 percent in 1983–87. Thus, the number of U.S.-trained Ph.D. economists securing their first job in the United States may be declining. What remains unclear is how many foreign-trained Ph.D. economists find their first job in the United States. We also do not know how frequently economists take jobs abroad later in their career or enter the United States in midcareer. The internationalization of economics in U.S. graduate schools is likely to continue, and this development will undoubtedly facilitate international mobility within the profession. The Commission will try to assess the relevance of these changes in the market environment for economists for the content and structure of graduate education in economics.

ROBERT M. SOLOW, MIT: It was probably inevitable that the Commission would conduct a survey to provide some kind of objective basis for its study. Spreadsheet programs want data the way lions used to want Christians. In fact, the survey results seem predictably bland. Graduate students of economics and their teachers have more or less chosen to do what they are doing. Economists tend to be obtuse about such choices. We pretend not to realize that many people slide gradually into whatever they are doing until they may feel locked in by events and pressures without ever having made a clear decision. If Thoreau was right that most men lead lives of quiet desperation, it is pretty unhelpful to say that they chose to do so. Nevertheless, graduate students and faculty usually have effective alternatives. They are probably pretty happy with their situations, or at least reasonably well adjusted to them. So it is no wonder that many of them give bland answers to the questions they were asked. Of course, there are indications of trouble in River City, and I want to talk about them. But the numbers do not suggest mass disaffection.

Even the obvious dissonant note seems to have a straightforward explanation. Lee Hansen reports that the skills that Ph.D.s in economics need and use in their jobs are rather different from the skills emphasized in their courses. The Ph.D. programs emphasize the teaching of theory and econometrics. They aim at developing mainly analytical and mathematical skills. Life seems to call much less for mathematics and even a bit less for analytics and much more for creativity, critical judgement, and the ability to communicate. This looks like an error to be corrected. But, of course, the right thing to teach depends not only on its usefulness to the student, but also on the ability of the teacher to teach it. The obvious mathematical model leads trivially to the conclusion that optimal teaching must equalize across skills the product of the marginal significance of a skill, and the marginal productivity of effort in imparting it. (I was able to throw that conclusion into an elasticity form that made it look a little more mysterious. It took great moral courage to refrain from writing it up and submitting it to a learned journal.)

I suspect that the reason we emphasize theory and econometrics and focus on developing our students' analytical and mathematical skills is because that is what we know how to do reasonably well. If I knew how to teach creativity and critical judgement, I guess I would do it. I know how to recognize them all right, and can even offer examples. But, that is far from being able to teach them. Communication skills are much more straightforward, and I think we should emphasize them more. The trouble is that teaching communication skills is not our skill. But I do think I will go home and urge that we force all of our first-year graduate students into a writing course to be taught by the MIT Writing Program. It is well to remember that we get our graduate students at the age of 22, more or less, after they have been through 16 years of formal education. It is unlikely that anything we can do in the next couple of years will have great effect on their creativity or communication skills.

My guess is that those who feel that something needs to be done about graduate education in economics are really expressing a feeling that something is wrong with economics. That almost has to be the case, because graduate teaching in economics more or less is economics, at least in the 15 or 20 major graduate departments. We teach our students to be like ourselves. This could explain the apparently unfocused desire for "change" that appears in many, though far from all, of the responses. Graduate education may be doing a good job of supporting the current orientation of the discipline, but some of the teachers and students have their doubts about the discipline itself.

I think I find a clue to one of our problems in the well-known fact, documented in the survey, that Ph.D. students have a hard time finding a thesis topic, take much too long to find one, and are then inadequately prepared to do thesis research. Why is that? Could it be that we tend to overvalue a kind of spurious "originality" at the expense of "normal science?" I think so, but I am not sure. If it happens, it might be another consequence of the fact that economics is not an experimental science. So the option of doing the next clever experiment to refine a hypothesis or check a conjecture is not available. It is getting less plausible to mine the same old data one more time. That seems to leave no option for the student who wants to make a mark except to invent a brand-new idea and strike out in untraveled directions. That could be bad for students, who find themselves lost for too long, and bad for the discipline that does not advance in any given direction as far as it might. Many of us have the feeling that economics is not as cumulative as we would like. I would not want to imitate the situation in chemistry or biology where, according to gossip, graduate students are in large part mere laboratory assistants carrying out the master's research program. But a little bit of that might not hurt if we could figure out how to do it.

When the Commission comes to write its report, I hope they will consult themselves more than the survey. It is an awesome committee. (A certain U.S. Senator once defended a Nixon nominee to the Supreme Court by saying that the country was full of mediocre people—Senator Hruska would have known—and they deserved representation on the Court too. This committee is all Pro Bowlers. They need to hear from the third and fourth tiers.) I think we would rather know what they think about the state of the discipline and of its main institution for recruiting, training, and socializing new members than to have them work over the survey. I also hope the Commission does not try too hard to achieve wide consensus, which tends to be vapid. This is the sort of subject on which difference of opinion is not merely inevitable, but healthy. One or two strong recommendations might get our attention.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
HUNDRED AND SECOND
ANNUAL
MEETING

ATLANTA, GEORGIA
DECEMBER 28–30, 1989

The John Bates Clark Award

Citation on the Occasion of the Presentation
of the Medal to

DAVID KREPS

December 29, 1989

David Kreps's contributions to economic research are nothing short of pathbreaking. This is evident in finance, where his work on arbitrage pricing is especially noteworthy, and in his modeling of rational expectations equilibria. Also important are Kreps's extensions of utility theory to dynamic contexts (temporal resolution of uncertainty), induced preferences, and flexibility. Especially and insistently, however, he has made pioneering contributions to dynamic game theory. Kreps's research goes to core issues. Rather than finesse "complicated" behavioral attributes of human actors by employing behavioral assumptions that are accommodated by state-of-the-art apparatus, Kreps insists that the complications be faced instead. Issues of learning, reputation, culture, unforeseeability, and the like that were once believed to be beyond reach are brought within the ambit. Kreps combines keen intuition, deep curiosity, and extraordinary skills of modeling and exposition. Innovative and tractable new models are his trademark. The added breadth and depth of his contributions have brought new life to and renewed excitement in the study of economics.

Minutes of the Annual Meeting
Atlanta, Georgia
December 29, 1989

The one hundred and second annual meeting of the American Economic Association was called to order by President-elect Gerard Debreu at 6:30 P.M., December 29, 1989, in the Marquis Ballroom of the Atlanta Marriott Marquis Hotel, Atlanta, Georgia. Debreu opened the meeting by explaining that the untimely death of Joe Pechman, who would have chaired the meeting, was the reason that he was conducting the meeting. He then reviewed the agenda and the order of business.

Debreu called for the approval of the minutes of the last annual meeting as published in the *Papers and Proceedings* issue of the *American Economic Review* (May 1989, p. 391). Carl Lundgren moved and Don Agthe seconded that the minutes be corrected as follows: After the sentence that begins, "Eisner ruled the resolution out of order..." add the following sentences: "Eisner also stated that the Association cannot appoint a committee to do what the Association itself cannot do. Eisner also cited the Executive Committee's prior vote (on December 27, 1988) as further support for his decision to rule the resolution out of order. (Lundgren was neither invited nor permitted to attend this meeting of the Executive Committee.)"

After the sentence that reads, "Lundgren opposed the ruling, arguing that publication of articles in the Association's journals was analogous to what he wanted for his 'invention.'" Change the period after "invention" to a comma and add the following words to extend the sentence: "noting that the journals are permitted to publish partisan opinions on economic subjects."

The Secretary responded that he did not object to the second part of the motion, but he did not accept the correction offered in the first part. The parenthetical statement concerning Lundgren's absence from the deliberations of the Executive Committee meeting was already reflected in the minutes of that meeting, which are also published in the *American Economic Review*. The names of persons attending are contained therein. The offered correction was unnecessary and inappropriate. Regarding the part of the proffered correction to the remarks of Robert Eisner, the Secretary stated that the minutes were not a verbatim transcript of the meetings. He had written the minutes immediately after the meeting from his notes. He believed they accurately reflected the substance of Eisner's remarks. Eisner stated that he preferred to accept the Secretary's version of what took place rather than attempt to reconstruct what might have been said one year ago.

Leo Raskind moved, and it was seconded, the previous motion. Debreu informed the assembly that the motion was not debatable and required a two-thirds vote to pass. The motion to close debate was passed. The meeting then voted on Lundgren's original motion to correct the minutes as proposed. The motion failed. Debreu ruled that the minutes stood as published.

The Secretary (C. Elton Hinshaw), Treasurer (Hinshaw), Editor of the *American Economic Review* (Orley Ashenfelter), Editor of the *Journal of Economic Literature* (John Pencavel), Managing Editor of the *Journal of Economic Perspectives* (Timothy Taylor), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports (published elsewhere in this issue). These had been distributed to members prior to the meeting. Lundgren asked the Director of *JOE* if it was true that one had to be a member of the Association to subscribe to *JOE*. Hinshaw responded that it was indeed true. *JOE* had been initiated as a service to AEA members, and the Association subsidized its publication. Lundgren opined that the constraint hit those least likely to be able to afford both membership dues and a subscription to *JOE*—the unemployed. He thought that some other financial arrangement should be instituted.

Upon the completion of the above listed reports, the Chair called for further business. Frank Thompson moved and Robert Pollin Seconded the resolution that had been submitted in timely fashion prior to the meeting, and had been made available to the members along with the agenda and reports. The resolution read as follows:

> Let it hereby be resolved that: The American Economic Association will not in the future rent or lease facilities wholly or partly owned by the John Portman Companies of Atlanta, Georgia, as long as the John Portman Companies remain on the published National Boycott List of the American Federation of Labor-Congress of Industrial Organizations.

Elaine McCrate spoke in favor of the resolution. She said that because John Portman Properties was on the AFL-CIO boycott list for alleged unfair labor practices, many members, particularly those with strong ties to the labor movement, felt very uncomfortable about meeting in a hotel in which Portman Properties owned an equity interest. She added that the current labor dispute between the Service Employees International Union and Portman Properties stems from an ongoing attempt by the SEIU to organize the janitors. The union is seeking higher wages and better working conditions. It alleges that Portman Properties has harassed members and has subcontracted the cleaning services in order to circumvent the bargaining process. The union has filed unfair labor practice charges with the National Labor Relations Board.

Leo Raskind, Legal Counsel for the Association, gave some background information about what had transpired in Atlanta this year and argued that the resolution was not needed to accommodate the site preferences of the members. The Association entered into contracts with the Atlanta hotels about five years ago. If a labor dispute existed at that time, no one was aware of it. Indeed, it was not until October of this year that one of our affiliated societies (the Industrial Relations Research Association) became aware of the AFL-CIO boycott of the Westin Peachtree Plaza and the Atlanta Marriott Marquis. All of this organization's sessions had been scheduled in the Westin Peachtree Plaza, a hotel owned by Portman Properties, but managed by the Westin Hotel Corporation. When the AEA Secretary was informed of the problem and how sensitive the IRRA members are to labor relations issues, he succeeded in moving their entire conference, including sessions, business and meal functions, and sleeping rooms to another hotel in the ASSA block of hotels. He also succeeded in moving two or three of the sessions organized by the Union of Radical Political Economists. In summary, the Association had shown itself to be remarkably sensitive to these issues and had done everything possible to accommodate the preferences of the societies involved. He further noted that both sides to the dispute have filed unfair labor practice charges with the NLRB and that no determinations have been made. The issues are currently under consideration by the NLRB. The resolution calls for the Association to take a stand on an issue that is currently in litigation.

David Gordon spoke in favor of the resolution. He argued that the resolution did not call for the Association as such to take a stand on a partisan issue, but directly involved the business affairs and activity of the Association. As with the issue of South Africa divestment, the AEA is involved in this dispute simply by virtue of the disposition of its money. Those sponsoring the resolution would prefer that the Association not side with Portman Properties by favoring them with its patronage.

Robert Eisner spoke against the resolution. The Association had demonstrated that it was sensitive to these issues and had tried to resolve the problems caused by it when made aware of the members' concerns. He felt that the resolution did call for the Association to take a partisan stand, and that that was the wrong way to proceed. Given the diversity of views of AEA members, the Association should not take positions on such issues. The Chair called for the vote. The resolution passed, 27 to 12. Debreu reminded the assembly of provision V.2 in the bylaws that allows the Executive Committee to submit a resolution adopted at an annual meeting in which less that 5 percent of the membership of the Association has voted to a mail ballot.

Debreu then recognized the new President-elect, Thomas Schelling. There being no further business, the meeting was adjourned at 7:15 P.M.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in Washington, D.C., March 24, 1989.

The first meeting of the 1989 Executive Committee was called to order at 10:00 a.m., March 24, 1989, in the Sheraton Washington Hotel. Members present were Joseph Pechman (presiding), G. A. Akerlof, Orley Ashenfelter, Gary Becker, Andrew Brimmer, Gerard Debreu, Robert Eisner, Stanley Fischer, C. Elton Hinshaw, John Pencavel, Isabel Sawhill, Joseph Stiglitz, Lawrence Summers, and Judith Thornton. Present for part of the meeting were members of the Nominating Committee—Alice Rivlin, Chair, Vernon Dixon, David Fand, David M. Gordon, Eric Hanushek, Irene Lurie, and Janet Yellen; and the Committee on Honors and Awards—Dale W. Jorgenson, acting chair, James J. Heckman, and John B. Taylor. Present for purposes of giving a report was William Baumol. Leo Raskind was present as Counsel.

Pechman opened the meeting by reviewing the agenda. He then asked for corrections or approval of the minutes of the December 27, 1988, meeting. After the correction of the first name of one of those present at the last meeting, the minutes were approved as corrected.

Ad Hoc Committee on the Journal of Economic Perspectives (Baumol). Baumol, who chaired the committee, briefly reviewed the Committee's written report that had been circulated prior to the meeting. (The full report is printed elsewhere in this issue.) The ensuing discussion centered on two issues raised by the report—"unbundling" and "open" submissions. After an extended discussion, it was VOTED to undertake a three-year unbundling experiment along the following lines. Members would be offered the option of not receiving any one of the three journals, with a reduction in membership dues approximately equal to the average of the marginal costs of the journals.

Under the current policy, articles for the *Journal of Economic Perspectives* are commissioned and solicited by the editors. Many members resent the policy. The *Ad Hoc* Committee was sensitive to the complaints but, on balance, favored continuation of the present policy and suggested some steps that could be taken to diminish the perception of some that the journal was the instrument of "the establishment." Given the evidence of general member satisfaction with the contents of the journal, the AEA's portfolio of journals, the cost of administering an open submissions policy, the large number of existing journals that are refereed, and the probability that a change to an open submissions policy would change the nature and character of the journal, it was decided to continue the present policy. It was understood that the *JEP* editor would continue to endeavor to insure that the contents of the journal are weighted heavily toward the preferences of the general membership rather than "the establishment," that is, the editor should continue to pursue the goal for which the journal was founded.

Report of the Editor of the American Economic Review (Ashenfelter). As his only item of business, Ashenfelter asked that John Kennan's term on the *AER*'s Board of Editors be extended by one year. It was VOTED to accept his recommendation.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel reported that he was currently negotiating with a firm to make the bibliographic material in the *JEL* available on compact disc in addition to the Indexes and Dialog. He also raised the possibility of returning to the former policy of paying authors for articles that the *JEL* commissioned. He was beginning to realize some competition for review-type articles from the *JEP*. It was generally conceded that review articles for the *JEL* were more difficult and time consuming to write and that payment might be necessary. No decision was made.

Report of the Editor of the Journal of Economic Perspectives (Stiglitz). Stiglitz distributed the tentative Table of Contents for the next few issues of the journal. He anticipated publishing symposia on budget deficits, auctions, real business cycles, bank deposit insurance, and LDC debt along with articles and the standard features—puzzles, further reading, and anomalies.

Report of the Secretary-Treasurer (Hinshaw). The Secretary reported that this year's annual meeting would be held in Atlanta (December 28–30) and next year's in Washington, D.C. Anaheim, Chicago, and New Orleans are being considered as sites for the January 3–5, 1992 meeting.

Registration for the 1988 New York meeting totaled 8,203. Thirty-nine other associations met with the AEA, 455 sessions were organized, and 109 events (luncheons, cocktail parties, committee meetings, etc.) scheduled. Each of these numbers is a record high for the meetings.

A new survey of members will be published in December. Members will receive biographical questionnaires this spring. In keeping with the Executive Committee's desire to keep "directory" information as current as possible, the Secretary proposed that the present style directory be published less often and that a "telephone and address" directory be published every other year. It was decided to move to such an arrangement —a four-year cycle for the "full fledged" directory and a two-year cycle for the "telephone and address" one.

It was VOTED to approve the following change in the bylaws. The bylaw to be changed is Article I, Section 2. It now reads:

> There shall be six classes of members other than honorary: regular members with annual incomes of $30,000 or less paying the base fee defined below; regular members with annual incomes above $30,000 but not more than $40,000 paying one and one-fifth times the base fee; regular members with annual incomes above $40,000 paying one and two-fifths times the base fee; family members (persons living at the same address, as a regular member, additional memberships without subscription to the publications of the Association) paying one-fifth of the base fee; junior members (available to registered students for three years only) paying one-half the base fee; and life members comprising those who qualified for life membership by making a single payment of the designated amount prior to January 1, 1976, and exempt from annual fees.
>
> Effective January 1, 1976, the base fee is $25.00 per year. The Executive Committee may increase the base fee in proportion to the increase occurring after January 1, 1976 in relevant price and wage indexes. It may increase the income brackets for regular members but may not decrease them below the figures specified in this bylaw.

The change eliminates the words "for three years" from the second parenthetical phrase —"available to registered students for three years only." It was understood that the Executive Committee would, if empowered to do so, set the limit at 5 years. The new bylaw would read:

> There shall be six classes of members other than honorary: regular members with annual incomes of $30,000 or less paying the base fee defined below; regular members with annual incomes above $30,000 but not more than $40,000 paying one and one-fifth times the base fee; regular members with annual incomes above $40,000 paying one and two-fifths times the base fee; family members (persons living at the same address, as a regular member, additional memberships without subscription to the publications of the Association) paying one-fifth of the base fee; junior members (available to registered students only) paying one-half the base fee; and life members comprising those who qualified for life membership by making a single payment of the designated amount prior to January 1, 1976, and exempt from annual fees.
>
> Effective January 1, 1976, the base fee is $25.00 per year. The Executive Committee may increase the base fee in proportion to the increase occurring after January 1, 1976 in relevant price and wage indexes. It may increase the income brackets for regular members but may not decrease them below the figures specified in this bylaw.

The Treasurer distributed the audited financial statements for 1987 and 1988. The "Statements of Revenues and Expenses" show an operating deficit of $398 thousand for 1988, up from $238 thousand the previous year. After taking into account investment gains of $281 thousand in 1988, there was a deficit of $117 thousand. The actual operating loss is smaller than budgeted ($510 thousand) and the total deficit compares to a budgeted deficit of $247 thousand.

The ratio of net worth as of December 31, 1988, to budgeted expenditures for 1989 is 1.41. A ratio of 1.0 is considered ample for safety purposes. The net worth at the end of 1988 was $3,966 thousand. The budget adopted by the Executive Committee at its December 27, 1988, meeting calls for expenditures of $2,808 this year. Thus, the excess over safety requirements of one-year's expenditures is on the order of $1.2 million.

The budget adopted for 1989 showed an operating loss of $471 thousand, investment income of $265 thousand, and an overall deficit of $206 thousand. Now that the audited figures for the end of 1988 are available, the investment income for the 1989 year is known, viz., $297 thousand, decreasing the projected deficit to $174 thousand.

Committee on Honors and Awards (Jorgensen). Jorgenson, acting as chair of the Committee, reviewed the procedures followed to select candidates for the John Bates Clark Medal. The Electoral College, consisting of the reporting committee and the Executive Committee VOTED to award the 1989 Clark Medal to David Kreps. Again acting as the Electoral College, they VOTED to elect Jean-Michel Grandmont, Angus Madison, Takashi Negishi, and Kazushi Ohkawa as Foreign Honorary Members.

Nominating Committee (Rivlin). The Electoral College, comprising the Nominating and Executive Committees, chose Thomas Schelling as the nominee for President-elect and Jacob Mincer and Guy Orcutt as Distinguished Fellows. Rivlin reported the Nominating Committee's choices of candidates for other offices: for Vice-President (two to be elected) Rudiger Dornbusch, Allan H. Meltzer, Walter Oi, and Albert Rees; for members of the Executive Committee (two to be chosen), Donald Brown, Gregory C. Chow, Richard R. Nelson, and Susan Rose-Ackerman.

Other Business (Pechman). It was VOTED to renew the terms of Stiglitz, Editor of the *JEP*, Raskind, Counsel, and Hinshaw, Secretary-Treasurer, for another three years. Their terms of office will now end December 31, 1993.

It was VOTED to participate with the Association for the Advancement of Science and the American Council of Learned Societies in a program to distribute scholarly journals to institutions in sub-Saharan Africa. It was also VOTED to authorize the President to establish a committee to consider whether persons from other low-income countries and countries with exchange controls should receive special treatment in the structure of dues and subscriptions.

1989 Program (Debreu). The President-elect related that the 1989 AEA program was virtually complete and distributed a preliminary program of sessions. He expected very few changes. Of the 89 sessions organized, 33 were "invited," 26 "contributed," and 24 joint with sister organizations; 6 sessions were organized by CSWEP. Two hundred-fifty-two papers had been submitted. They were very unevenly distributed among the 10 fields. He also announced that Michael Boskin would give the AEA/AFA Joint Luncheon address, Janet Norwood will give the SGE Distinguished Lecture, and David Landes, the Ely Lecture.

There being no additional business, the meeting adjourned.

Minutes of the Meeting of the Executive Committee in Atlanta, Georgia, December 27, 1989,

The second meeting of the 1989 Executive Committee was called to order at 10:10 A.M., December 27, 1989, in the Chablis Room of the Atlanta Marriott Marquis Hotel in Atlanta, Georgia. Members present were Gerard Debreu (presiding), George Akerlof, Orley Ashenfelter, Robert Barro, Alan Blinder, Andrew Brimmer, Robert Eisner, Stanley Fischer, C. Elton Hinshaw, John Pencavel, Isabel Sawhill, Lawrence Summers, and Judith Thornton. Present as guests as newly elected members of the 1990 Executive Committee were Gregory Chow, Allan Meltzer, and Thomas Schelling. Present for parts of the meeting to give reports were Nancy Gordon, Anne Krueger, and W. Lee Hansen. Leo Raskind was present as Counsel.

(Joe Pechman, President of the Association for 1989, died on August 19, 1989. The AEA's bylaws (Article III, Section 1) provide that "his duties shall devolve upon the President-elect..." President-elect Debreu assumed the duties of President and chaired the meeting.)

Debreu opened the meeting by informing the group of the arrangements that had been made to commemorate Pechman. After recognition of the new Distinguished Fellows and awarding of the Clark Medal and immediately prior to the Presidential Address session, George Break, Alicia Munnell, and John Shoven would give brief eulogies. Mrs. Pechman would be recognized. Henry Aaron would read Pechman's speech.

Minutes. The minutes of the meeting of March 24, 1989, were approved as written. They were circulated to the members of the Committee prior to the meeting.

Report of the Secretary (Hinshaw). The Secretary welcomed the new members of the 1990 Executive Committee who were present (Schelling, Chow, and Meltzer) and thanked those whose terms were expiring (Blinder, Brimmer, Barro, Thornton, and Becker). The 1990 annual meeting will be held in Washington, D.C., on December 28–30, the last meeting at the traditional dates between Christmas and New Year's. The Executive Committee will meet on the 27th. In accordance with the results of the preferential vote of the members concerning the dates of future meetings, the next meeting will be moved to January 3–5, 1992. The Executive Committee will meet on Thursday, January 2. It was VOTED to approve the Secretary's recommendation of New Orleans as the site for the 1992 meeting, and he was authorized to negotiate contracts with the appropriate hotels in that city. San Francisco and Anaheim were discussed as possible sites for the 1993 meeting. No decision was made. Some preferred San Francisco. Others thought we should "spread our wings" a little and try a different type of city such as Anaheim.

The 1989 Survey of Members was mailed to the members of the Association early this December. It contains, among other things, biographical sketches of 22,219 persons. The Secretary reported that he was exploring ways to produce a telephone-book-type directory on a two-year cycle. It would contain name, address, phone and fax numbers, but no biographical material. The discussion of this new directory indicated that it might be feasible to send it to all members without a separate charge. Again, no decision was made. Debreu indicated that both the "new" and the "old" directories should be on four-year cycles so that one of them appears every other year—a two-year cycle for a directory.

The Secretary reported on two commercial "opportunities" available to the Association. Trans National Financial Services had proposed an affinity credit card program (the card would carry the AEA name and its logo). The Association would receive a fee which would be negotiated. It was VOTED to pursue the matter further. Albert Wohlers & Co., an insurance broker and administrator, had proposed a long-term health care plan (nursing home and home health care) for members. Discussion of the proposal indicated that the Secretary should not pursue the idea.

In order to meet postal regulations concerning second-class mailing privileges, it was VOTED to adopt the following resolution:

> Be it resolved that membership in the American Economic Association entitles one to receive the *American Economic Review*, the *Journal of Economic Literature*, and the *Journal of Economic Perspectives*. A copy of each issue of these journals shall be sent to each member unless the member requests otherwise. Of each member's annual dues, 90 percent shall be for subscriptions to the publications: 30 percent shall be for a year's subscription to the *American Economic Review*, 30 percent for a year's subscription to the *Journal of Economic Literature*, and 30 percent for a year's subscription to the *Journal of Economic Perspectives*.

The next meeting of the Executive Committee will take place March 23 (Friday) in Washington, D.C.

Report of the Editor of the American Economic Review (Ashenfelter). The Editor briefly reviewed his written report which is published elsewhere in this issue. It was VOTED to approve his recommendation of Paul Milgrom to replace Hal Varian as a Co-editor. It was also VOTED to approve his recommendations of the persons whose names are listed below as members of the Board of Editors: Reuben Gronau, Daniel Hamermesh, Kevin Hoover, Thomas Romer, Kenneth West, and Leslie Young.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel briefly reviewed his written report which was circulated prior to the meeting and is published elsewhere in this issue. It was VOTED to approve his recommendation of Charles Brown and John Riley as new members of the Board of Editors and to reappoint Duncan Foley. The revision of the journal's bibliographic classification scheme is proceeding. The new system will be expanded to include at least 19 basic and reorganized categories compared to the current 11. He hopes to introduce and operate the new system in 1990.

Report of the Editor of the Journal of Economic Perspectives (Stiglitz). Stiglitz reviewed his written report which is published elsewhere in this issue and was available prior to the meeting. It was VOTED to approve his recommendations to the Board of Editors—James Heckman, Barry Nalebuff, and Sherwin Rosen.

Report of the Director of Job Openings for Economists (Hinshaw). Hinshaw reviewed his written report which is published elsewhere in this issue. He pointed out that a record number (2,014) of new jobs and had been advertised during 1989.

Report of the Committee on the Status of Women in the Economics Profession (Gordon). Gordon reviewed her written report which had been circulated to the Executive Committee prior to the meeting and is published elsewhere in this issue. Discussion of the report centered on the slow progress of women through the professional ranks. There has been adequate growth in the number of women in the profession during the past decade, but this increase is not spreading to the higher, tenured ranks as yet.

The 1990 AEA Program (Schelling). Schelling, President-elect for 1990, reported that he had appointed a program committee of 16 persons, one for each category of his own revision of the *JEL* classification scheme. He was pleased to announce that George Akerlof had agreed to give the Ely Lecture. The program would concentrate on the revolutionary transformation from planned to market-oriented economies currently taking place, and a secondary emphasis would be on Western Europe in 1992.

Other Business. A resolution for the annual business meeting of the Association submitted by Frank Thompson was discussed. (See the Minutes of the Annual Business Meeting for the text of the resolution.) Opinions varied from "pernicious and in violation of our Articles of Incorporation" to "harmless and shouldn't be taken too seriously." During the course of the discussion, it was VOTED to recommend a change in the bylaws that would allow a mail balloting of the membership at the "next regular mailing" rather than within the 90 days which is currently required (Article V, Section 2). Mailings to the members cost approximately $10,000. If a special mailing could be piggybacked on a regular mailing, the decision about whether to ballot the membership on controversial resolutions passed at the annual meeting would be much easier. The Secretary was instructed to prepare the proposed amendment. The Association had been invited to participate (by giving money) in a demonstration project, now underway in the Washington Metropolitan area, that matches skills and knowledge of retired professionals with volunteer opportunities. The project is based on a similar one begun by the American Association of Retired Persons and now operated under the auspices of the AAAS. It was decided not to contribute.

Report of the Commission on Graduate Education (Krueger and Hansen). Krueger, who chairs the Commission, and Hansen, its Executive Secretary, reported that the Commission had about finished its work. She expected that its recommendations would be available by the end of January and a general report with specific studies of various aspects of graduate training would be ready by the end of the current academic year. Although the Commission will not speak for the Association, it was created and approved by the AEA. The issue now is to what extent does the AEA wish to participate in the dissemination of the Commission's report. As she now perceives it, there will be a short report on its recommendations—similar to an extended "executive summary"—and a long report of monograph length—the general report with specific studies. It was VOTED to approve publication of the summary report with its recommendations in one of the AEA's journals and support the independent publication of the larger work, separate from the AEA. It was understood that Krueger and the three editors would mutually determine which of the journals was the most appropriate one for publication of the short version.

Report of the Treasurer (Hinshaw). Hinshaw reviewed his report which is published elsewhere in this issue. It was VOTED to approve the 1990 budget as submitted to the Budget Committee and approved by it subject to the following understandings: (1) at the March meeting, the Treasurer would propose a rate increase effective January 1, 1991; (2) the policy of increasing nominal dues less than the rate of inflation, allowing real dues to decline, would have to be abandoned; (3) some "catching-up" in real dues was in order; (4) future Executive Committees should be very cautious about initiating new projects; and (5) existing programs should be reviewed carefully by the appropriate person with an eye toward containing cost increases. During the discussion of the report, several members indicated that the Association should strive to eliminate the deficit by a specific date. The date most often mentioned was 1993.

There being no further business, the meeting adjourned at 3:50 P.M.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*,

Report of the Secretary for 1989

Annual Meeting. The annual meeting will be held in Washington, D.C. in 1990. The meeting is scheduled for December 28–30 and will have a Placement Service, which will open for business one day earlier (December 27) than the meetings.

Elections. In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Alice M. Rivlin, Chair, Vernon J. Dixon, David I. Fand, David M. Gordon, Eric A. Hanushek, Irene Lurie, and Janet L. Yellen submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-elect
Thomas C. Schelling

Vice-President	*Executive Committee*
Rudiger W. Dornbusch	Donald J. Brown
Allan H. Meltzer	Gregory C. Chow
Walter Y. Oi	Richard R. Nelson
Albert E. Rees	Susan Rose-Ackerman

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of ballots, I certify that the following persons have been duly elected to the respective offices:

President-elect (for a term of one year)
Thomas C. Schelling
Vice-Presidents (for a term of one year)
Rudiger W. Dornbusch
Allan H. Meltzer
Executive Committee (for a term of three years)
Gregory C. Chow
Susan Rose-Ackerman

In addition, I have the following information:

Number of legal ballots	4,869
Number of invalid envelopes	270
Number of envelopes received after October 1	12
Number of envelopes returned	5,151

The proposed amendment to the bylaws allowing the Executive Committee to determine the number of years that one can be a junior member was adopted.

Membership. The total number of members and subscribers is shown in Table 1. The total has fluctuated between 25,000 and 26,500 since 1975. This year it reached an all-time high of 27,306.

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of Year)

	1987	1988	1989
Class of Membership			
Annual	17,230	17,410	18,172
Junior	1,549	1,939	2,170
Life	357	353	338
Honorary	31	33	33
Family	448	437	470
Complimentary	477	475	387
Total Members	20,092	20,647	21,570
Subscribers	5,748	5,793	5,736
Total Members and and Subscribers	25,840	26,440	27,306

National Registry. The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision at the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their openings.

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review*, *Journal of Economic Literature* and the *Journal of Economic Perspectives* totaled 696 in 1989, compared to 436 in 1988. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

AEA Staff. Mary Winer, Kimberly Adair, Norma Ayres, Dana Coleman, Sherry Davis, Violet Sikes, and Jacquelyn Woods are extraordinarily effective in handling the daily operations of the Association. Marlene Hall organizes the annual meeting efficiently and with a minimum of fuss. They are the Secretariat for all practical purposes. Their "front man" thanks them for their dedication and loyalty.

Committees and Representatives. Listed below are those who served the Association during 1989 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they were appointed. On behalf of the Association, I thank them for all their service.

Ad Hoc Committee to Evaluate the Journal of Economic Perspectives
William J. Baumol, Chair
Leonid Hurwicz
Michael McPherson
Albert E. Rees
Robert M. Solow

Budget Committee
C. Elton Hinshaw, Chair
Judith Thornton (1989)
Isabel V. Sawhill (1990)
Lawrence H. Summers (1991)
Joseph A. Pechman (1989)
Gerard Debreu (1991)

Census Advisory Committee
Ben E. Laden (1990), Chair
Ben S. Bernanke (1989)
Dennis W. Carlton (1989)
Dale T. Mortensen (1990)
Victor Zarnowitz (1990)
Timothy F. Bresnahan (1991)
Michael McKelvey (1991)
Margaret C. Simms (1991)
Thomas W. Synnott III (1991)

Committee on Economic Education
John Siegfried, Chair (1990)
William B. Walstad (1989)
Michael J. Boskin (1990)
Robert J. Highsmith (1990)
Robert T. Michaels (1990)
Robin L. Bartlett (1991)
Donald N. McCloskey (1991)
Phillip Saunders, Jr. (1991)
William E. Becker, Jr., *ex officio*

Economics Institute Policy and Advisory Board
Edwin S. Mills, Chair (1990)
Samuel A. Morley (1990)
Koichi Hamada (1991)
Ray Marshall (1991)
Daniel W. Bromley (1992)
James A. Millar (1992)
John P. Evans (1993)
Lawrence J. Lau (1993)

Finance Committee
C. Elton Hinshaw, Chair
Robert J. Genetski (1989)
Robert Hamada (1990)
B. Douglas Bernheim (1991)

Commission on Graduate Education in Economics
Anne O. Kreuger, Chair
W. Lee Hansen, Executive Director
Kenneth J. Arrow
Olivier J. Blanchard
Alan S. Blinder
Claudia Goldin
Edward E. Leamer
Robert J. Lucas
John C. Panzar

Rudolph G. Penner
T. Paul Schultz
Joseph E. Stiglitz
Lawrence H. Summers

Committee on Honors and Awards
Oliver E. Williamson, Chair (1991)
James J. Heckman (1989)
Richard R. Nelson (1989)
Dale W. Jorgenson (1991)
Alan S. Blinder (1993)
Hollis Chenery (1993)
John B. Taylor (1993)

1989 Nominating Committee
Alice M. Rivlin, Chair
Vernon J. Dixon
David I. Fand
David M. Gordon
Eric A. Hanushek
Irene Lurie
Janet L. Yellen

Committee on Political Discrimination
Burton A. Weisbrod, Chair (1989)
Stephen E. Margolis (1989)
Benjamin J. Cohen (1990)
Clark W. Reynolds (1990)

Committee on the Status of Minority Groups in the Economics Profession
Ronald L. Oaxaca, Chair (1991)
William A. Darity, Jr. (1990)
William D. Bradford (1990)
Susan M. Collins (1990)
George J. Borjas (1991)
Vernon J. Dixon (1991)
Clifford E. Reid (1991)
Margaret C. Simms (1991)

Committee on the Status of Women in the Economics Profession
Nancy M. Gordon, Chair (1990)
Cecilia Conrad (1989)
Shulamit Kahn (1989)
Judith R. Lave (1989)
Marjorie B. McElroy (1989)
Kathryn Morrison (1990)
Barbara Newell (1990)
Elizabeth Hoffman (1991)
Shelly J. Lundberg (1991)
June O'Neill (1991)
Daniel H. Newlon (1991)
Joseph A. Pechman, *ex officio*
Joan G. Haworth, Membership Secretary

Committee on U.S.-China Exchange
Gregory C. Chow, Chair
Kenneth Arrow
Lawrence R. Klein
Theodore W. Schultz

Committee on U.S.-Soviet Exchange
Lawrence R. Klein, Chair (1990)
Zvi Griliches (1990)
Edward A. Hewett (1990)
Kenneth Arrow (1991)
Robert Eisner (1991)
Franklyn D. Holzman (1991)
Leonid Hurwicz (1991)
Stanley Johnson (1991)
Richard Quandt (1991)
Nathan Rosenberg (1991)
Richard N. Rosett (1991)
Joseph Pechman, *ex officio*

COUNCIL AND OTHER REPRESENTATIVES

American Association for the Advancement of Science, Sec. K, Social Economics and Political Sciences
Adam Rose (1992)

American Association for the Advancement of Slavic Studies
Joseph Brada (1989)

American Council of Learned Societies
C. Elton Hinshaw (1990)

Consortium of Social Science Associations (COSSA)
Isabel V. Sawhill (1992)
C. Elton Hinshaw

Council of Professional Associations on Federal Statistics (COPAFS)
Charles R. Hulten (1991)
Joel Popkin (1989)

International Economic Association
Kenneth Arrow (1990)
C. Elton Hinshaw

National Bureau of Economic Research
David A. Kendrick (1990)

Policy Board of the Journal of Consumer Research
Louis L. Wilde (1991)

Social Science Research Council
Robert M. Coen (1990)

REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1989

Inaugurations
F. C. Richardson, State University of New York-College at Buffalo
Patricia A. Hutton

Report of the Treasurer for the Year Ending December 31, 1989

The proposed budget for 1990 in Table 1 projects an operating loss of $675 thousand, investment income of $296 thousand, and an overall deficit of $379 thousand. The recent and projected deficits result primarily from the decision to begin publication of the *Journal of Economic Perspectives*. Net worth at the beginning of 1990 is expected to exceed budgeted expenditures by more than a million dollars so immediate action is not required. Another deficit can be accommodated.

Audited statements for 1989 will be published in the June issue of the *American Economic Review*.

I wish to thank Norma Ayres, our accountant, and Mary Winer, the Administrative Director, for their valuable help and patience in assisting me in carrying out the duties of the Treasurer.

C. ELTON HINSHAW, *Treasurer*

TABLE 1—1990 BUDGET, AMERICAN ECONOMIC ASSOCIATION
(thousands of dollars)

	First Nine Months (Unaudited)		Actual	Full Year Budgeted	
	1988	1989	1988	1989	1990
REVENUES FROM DUES AND ACTIVITIES					
Membership dues	$700	$772	$944	$1,012	$1,041
Nonmember subscriptions	557	560	757	724	757
Subtotal	1,257	1,332	1,701	1,736	1,798
Subscriptions, *Job Openings for Economists*	22	25	34	34	34
Advertising	90	101	128	134	140
Sale of *Index of Economic Articles*	152	100	176	188	197
Sales of copies, republications, handbooks	34	35	44	45	45
Sale of mailing list	34	41	49	50	60
Annual meeting	71	83	71	50	50
Sundry	75	63	93	100	100
Total Operating Revenue	**1,735**	**1,780**	**2,296**	**2,337**	**2,424**
PUBLICATION EXPENSES					
American Economic Review	529	586	709	765	849
Journal of Economic Literature	654	702	911	950	1,065
Journal of Economic Perspectives	246	292	361	376	428
Survey of Members	52	52	70	70	70
Job Openings for Economists	35	39	59	55	59
Index of Economic Articles	80	40	96	90	100
Subtotal	1,596	1,711	2,206	2,306	2,571
OPERATING AND ADMINISTRATIVE EXPENSES					
General and Administrative	265	290	367	392	410
Committees	38	40	55	55	60
Support of other organizations	53	44	66	55	58
Subtotal	356	374	488	502	528
Total Expenses	**1,952**	**2,085**	**2,694**	**2,808**	**3,099**
OPERATING GAIN (LOSS)	(217)	(305)	(398)	(471)	(675)
INVESTMENT GAIN (LOSS)	210	223	281	265	296
SURPLUS (DEFICIT)	(7)	(82)	(117)	(206)	(379)

Report of the Finance Committee

The Finance Committee of the American Economic Association met at the Chicago Club, Chicago, Illinois, at noon on December 14, 1989. Present were Doug Bernheim, Robert Genetski, Robert Hamada (members of the Committee), and C. Elton Hinshaw (Chairman of the Committee and Secretary/Treasurer of the Association); Harvey Hirschhorn, Stephen McNally, and Robert McNeill (representing Stein Roe & Farnham, investment counsel for the Association).

In 1987, the Committee reviewed recommendations presented by the AEA Committee on Indexing Association Funds concerning the long-term allocation of the Association's investment assets. As a result of that recommendation and the subsequent deliberation of the Finance Committee, it was agreed that the Association's portfolio be comprised of a combination of the Wells Fargo South Africa Free Index Fund, Stein Roe & Farnham's specialty equity mutual funds, and a bond portion managed by Stein Roe & Farnham.

This restructuring took place at the end of June 1988. Investments were made in the Wells Fargo South Africa Restricted Equity Index B, as well as the Capital Opportunities, International Growth and Special Fund, which are all managed by Stein Roe & Farnham. The fixed-income portion of the portfolio was invested in short/medium-term U.S. government agency notes and SRF's longer-term taxable funds (Managed Bonds & Government Plus).

As a result of the aforementioned asset allocation strategy, the overall performance of the Association's fund now reflects the combined efforts of Wells Fargo and Stein Roe & Farnham.

With respect to the 1989 performance (through 30 November) of the Association's portfolio, the total return of the account including cash, bonds and SR&F equity funds was +17.4 percent. Focusing on a longer-term period, over the five years ending with November 1989, the total return for the account was +90.6 percent. In dollars, the total return has been over $2,500,000 over the last five years.

The Committee reviewed the current directive relating to equity allocation, and it was decided that the portfolio guidelines of 50 to 75 percent invested in equities should remain intact. The fixed income maturity restriction (stating a maximum permissible average maturity of ten years) was *removed* and left to the discretion of the advisor. In addition, it was decided that a minimum cash position of 5 percent (of the total portfolio) should be maintained at all times, as a liquidity reserve.

Members can obtain a list of the assets in the portfolio by writing the Treasurer.

C. ELTON HINSHAW, *Chair*

Report of the Managing Editor

American Economic Review

The editorial process has worked very smoothly during the past year despite some changes in personnel at the *Review* and a substantial increase in manuscript submissions. Our current submission rate (of nearly 1,000 papers per year) has, after a two-year hiatus, returned to the record level reached in 1986.

Editorial Process

As Table 1 indicates, the number of submissions has increased by about 12 percent over last year. Nevertheless, the chances that a submitted paper will eventually be published have remained constant from last year to this year. The fraction of submitted papers that we publish has now been at the same all time historical low of .12 for three years.

Table 2 indicates that we published more papers and pages of material in the *Review* in 1989 than in 1988. Although I and my co-editors have adopted a conscious policy of attempting to increase the number of major substantive articles we publish at the expense of shorter papers, comments, and replies, we have not been very successful at doing so during the last year.

Tables 3 and 4, when compared with the results for last year, indicate there has been very little change in the speed with which we handle manuscripts that are ultimately rejected. A major cause of the delay in handling manuscripts can be attributed, as I indicated in my Report last year, to the increased use of referees and to the slowness with which we receive reports. There is clear evidence that our new policy of making payments to referees has speeded up the referee-

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1970–89[a]

Year	Submitted	Published	Ratio Published-to-Submitted
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14
1984	921	138	.15
1985	952	128	.13
1986	987	123	.125
1987	843	99	.12
1988	844	100	.12
1989	946	116	.12

[a] The submissions reported for every year refer to the last two months of the previous year and the first ten months of the year reported.

TABLE 2—SUMMARY OF CONTENTS, 1988 AND 1989

	1988 Number	1988 Pages	1989 Number	1989 Pages
Articles	52	807	51	851
Shorter Papers, including Comments and Replies	48	324	65	427
Announcements and Notes Section		31		26
Index		9		10
Total		1171		1314

ing process. At the same time, however, I and my co-editors have again increased our use of the refereeing process. These offsetting factors have resulted in only a small reduction in the delay with which we handle manuscripts.

Table 5 indicates that there has been some change from last year in the speed with which accepted papers are processed and published in the *Review*. We have now built up a backlog of accepted papers, and this has naturally led to an increase in the lag between acceptance and publication. Building up our backlog has, on the whole, been desirable, since we had been operating with far too tight a printer's deadline in the last three years.

The subject matter distribution of papers published in the *Review* in 1988 and 1989 is contained in Table 6. It remains my impression that the distribution of published papers reflects fairly accurately the distribution of papers submitted.

TABLE 3—DISPOSITION OF MANUSCRIPTS, 1988 AND 1989

	July 1, 1987– June 30, 1988	July 1, 1988– June 30, 1989
Manuscripts Received	835	923
Completed Processing	494	568
Accepted	21	38
Rejected	473	530
Currently in Process	341	355

Finally, as I reported earlier in the year, we have been testing for the impact of double-blind refereeing on the editorial process at the *Review* by randomly assigning part of our submissions to a double-blind refereeing process. This experiment is continuing and a report of its results should be available in mid-1990.

Papers and Proceedings

The eleventh volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1989. For several years this task has been very capably handled by Harvey Rosen (Princeton University) and Wilma St. John. Harvey Rosen resigned from his position last year, and I would like to express my gratitude for his several years of outstanding work. This year the task was handled by Ronald Oaxaca (University of Arizona) and Wilma St. John. I am deeply indebted to both of them for the difficult work under extraordinarily tight deadlines that they have so capably performed.

Co-Editors and Board of Editors

There have been several major changes in the editorial personnel at the *Review* in the last year. Hal Varian resigned his position as co-editor of the *Review* at the end of 1989. I am deeply grateful for all the help and effort

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION, JULY 1, 1988–JUNE 30, 1989

Weeks to Rejection	Total Number of Manuscripts	Percent	No Outside Referees	1 Referee	2 Referees	3 or More Referees
0–4	37	0.05	29	8	0	0
5–6	35	0.05	9	21	5	0
7–8	63	0.09	3	38	22	0
9–10	78	0.11	0	40	36	2
11–12	89	0.13	1	27	56	5
13–14	72	0.10	1	22	45	4
15–16	59	0.08	0	22	35	2
17–21	98	0.14	0	33	59	6
22–26	66	0.09	0	22	35	9
27–30	32	0.05	2	4	20	6
31–35	25	0.04	1	3	19	2
36–52	39	0.05	0	11	17	11
52+	18	0.02	1	3	10	4
	711	100.00	47	254	359	51

TABLE 5—AVERAGE PUBLICATION LAGS BY JOURNAL ISSUE

Journal Issue	Number of Weeks Lag		
	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
March 1989	39	28	67
June 1989	41	28	69
September 1989	53	32	85
December 1989	53	39	92

TABLE 6—SUBJECT MATTER DISTRIBUTION OF PUBLISHED MANUSCRIPTS, 1989 AND 1989

	Published	
	1988	1989
General Economics and General Equilibrium Theory	6	19
Microeconomic Theory	8	6
Macroeconomic Theory	4	1
Welfare Theory and Social Choice	1	2
Economic History, History of Thought, Methodology	5	1
Economic Systems	1	0
Economic Growth, Development, Planning, Fluctuations	9	10
Economic Statistics and Quantitative Methods	1	5
Monetary and Financial Theory and Institutions	8	4
Fiscal Policy and Public Finance	7	14
International Economics	18	14
Administration, Business Finance	4	4
Industrial Organization	15	12
Agriculture, Natural Resources	0	1
Manpower, Labor Population	4	13
Welfare Programs, Consumer Economics, Urban and Regional Economics	9	10
Total	100	116

he provided over his two and one-half year period as co-editor. Although Hal Varian has been the very model of an efficient editor, he has also left a longer-term legacy at the *Review*. As some economists know, Varian has an interest in, and facility with, computer software that is legendary. (He typeset his own textbook!) As a result, he has helped to ease the editorial staff at the *Review* into the maze of editorial software with skill, and (most important) good humor.

Paul Milgrom (Stanford University) graciously consented to take Varian's place as co-editor and so I now edit the *Review* with his assistance as well as that of Robert Haveman and Bennett McCallum. I am deeply indebted to them for the conscientious effort they have expended over the last year.

The Board of Editors now consists of twenty-two members and I am indebted to them also for their efforts. Board members are selected to reflect the highest level of scholarship in the economics profession from the breadth of different fields represented in our submissions. More than fine scholarship is expected of a Board member, however. Board members are also selected because of their conscientiousness, good judgment, and professional reliability. When possible, we like to select Board members from those

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1989 *AER*

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,000	293	344	$63,957.54	$1,422.48	$65,380
May	28,000	431	464	84,024.07	2,762.34	86,786
June	28,000	326	352	68,558.89	1,742.15	70,301
September	28,000	335	384	75,314.72	1,618.89	76,933
December[a]	28,000	351	400	77,000.00	2,200.00	79,200
Annual Misc.[b]						12,000
Total		1,736	1,944			$390,600

[a] Estimated.

[b] Estimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

economists who have been especially helpful in the outside refereeing process.

Two members of the Board completed their terms during 1989: George E. Johnson and Leslie Young. I am most grateful to them and to the continuing members: George Akerlof, James Anderson, Timothy Bresnahan, Henry Farber, Marjorie Flavin, Robert Flood, Claudia Goldin, Jo Anna Gray, Kenneth Judd, John Kagel, John Kennan, Dale Mortensen, Maurice Obstfeld, Edgar Olsen, John Riley, Richard Roll, David Sappington, Kenneth Singleton, Robert Smith, Barbara Spencer, Richard Tresch, and John Wilson. I am also grateful for the assistance of many other associates who make it possible to edit and produce the *Review*. I am indebted to Claire Comiskey, our production editor, and to our office manager, Shirley Griesbaum, and our editorial assistant, Sandra Grant, for the fine work they have continued to perform over the past year. I would also like to thank the co-editors' secretaries: Joan Sample (Robert Haveman's office); Pat Niber (Bennett McCallum's office); and Carolyn Bartle (Hal Varian's office).

As always, the published version of this report contains the list of referees who have volunteered their services during 1989. We extend our deepest appreciation for the time and energy they have devoted to the advancement of our science.

H. Aaron
M. Abbott
A. B. Abel
J. M. Abowd
K. G. Abraham
M. Abramovitz
B. A. Abrams
M. Adler
A. R. Admati
P. Aghion
S. Ahmed
G. A. Akerlof
A. Alesina
B. Allen
F. Allen
S. G. Allen
J. Alm
J. G. Altonji
S. J. Ambler
Y. Amihud
J. E. Anderson
J. Andreoni
J. Angrist
J. J. Antel
R. Antle
E. Appelbaum
R. J. Arnott
W. B. Arthur
D. A. Aschauer
P. Asquith
A. Atkinson
A. Auerbach
L. Auernheimer
R. Ayanian
C. Azariadis
D. Backus
K. Bagwell
R. T. Baillie
M. N. Baily
S. Baiman
J. B. Baker
B. A. Balassa
R. Baldwin
N. S. Balke
L. Ball
C. Ballard
E. Baltensperger
J. S. Banks
D. P. Baron
R. J. Barro
R. Barsky
A. Bartel
J. R. Barth
R. L. Bartlett
Y. Barzel
L. J. Bassi
K. Basu
B. W. Bateman
R. Batina
R. C. Battalio
W. J. Baumol
M. Baxter
C. Beach
R. Becker
M. J. Beckmann
J. R. Behrman
M. T. Belongia
D. Benjamin
A. Bergson
T. Bergstrom
D. Berheim
J. Berkovec
E. Berkovitch
B. Bernanke
E. R. Berndt
S. T. Berry
T. Besley
H. Bester
J. Bhagwati
M. Bils
K. Binmore
J. Bishop
R. Bishop
C. Blackorby
O. J. Blanchard
R. M. Blank
D. Blau
F. Blau
M. Blaug
M. I. Blejer
A. S. Blinder
C. Bliss
G. Blomquist
D. E. Bloom
B. Bluestone
L. Blume
R. Boadway
N. E. Bockstael
Z. Bodie
J. N. Bodurtha, Jr.
R. Bolton
W. A. Bomberger
E. Bond
J. P. Bonin
G. Bonnano
M. Bordo
S. Borenstein
G. Borjas
J. Boschen
P. Bossaerts
J. Boughton
J. Bound
A. L. Bovenberg
H. P. Bowen
J. Boyd
R. S. Boyer
R. M. Bradburd
D. Bradford
R. Braeutigam
J. A. Brander
M. Braulke
G. Brennan
T. F. Bresnahan
A. A. Brewer
C. C. Brown
G. M. Brown, Jr.
J. N. Brown
E. K. Browning
M. J. Browning
N. Bruce
M. Bruno
H. J. Bruton
J. M. Buchanan
W. H. Buiter
J. Bulow
J. B. Burbidge
R. V. Burkhauser
O. Burt
G. Burtless
W. Butz

G. Cain
L. P. Cain
C. W. Calomiris
G. Calvo
C. F. Camerer
R. Cameron
S. Cameron
J. Y. Campbell
M. B. Canzoneri
A. S. Caplin
D. Card
P. S. Carlin
J. B. Carlson
D. W. Carlton
L. H. Carmichael
M. R. Carter
A. C. Case
J. P. Caskey
J. H. Cassing
S. G. Cecchetti
R. G. Chambers
L. K. C. Chan
W. W. Chang
H. Chao
D. Chaplin
S. Chatterjee
J. Chavas
L. K. Cheng
R. S. Chirinko
B. R. Chiswick
G. C. Chow
C. F. Christ
L. J. Christiano
C. Y. C. Chu
Y. Chu
C. Clague
G. Clark
P. K. Clark
T. A. Clark
D. Cleeton
C. T. Clotfelter
S. Coate
J. H. Cochrane
R. M. Coen
B. G. Colby
W. Cole
S. Collins
W. S. Comanor
J. Conlisk
D. A. Conn
J. Conrad
R. F. Conrad
G. Constantinides
P. J. Cook
T. Q. Cook
T. F. Cooley
R. W. Cooper
W. M. Corden
J. J. Cordes
P. J. Coughlin
P. N. Courant
F. Cowell
D. Cox
J. C. Cox
S. Craig
P. Cramton
R. Crandall
J. Crane
V. P. Crawford
K. Crocker
T. Crocker
A. Cukierman
R. E. Cumby
J. Currie
D. M. Cutler
J. Danthine
L. Danziger
P. Danzon
M. R. Darby
W. A. Darity, Jr.
M. M. Darrough
S. P. Das
J. Davanzo
M. David
C. Davidson
J. B. Davies
A. V. Deardorff
A. S. Deaton
P. DeGraba
O. De La Grandville
J. B. De Long
G. Demange
N. B. De Marchi
D. de Meza
H. Demsetz
W. den Haan
E. R. Denison
A. Denzau
W. Desvousges
W. Dewald
D. W. Diamond
P. Diamond
W. T. Dickens
F. X. Diebold
Y. Dirickx
A. K. Dixit
N. A. Doherty
K. M. Dominguez
D. Donaldson
J. B. Donaldson
M. P. Dooley
M. Dotsey
S. Dowrick
A. Drazen
R. Driskill
L. Dudley
J. A. Dunlevy
D. S. Dutton
L. Dworin
G. P. Dwyer
P. Dybvig
R. Dye
J. Eaton
Z. Eckstein
N. Economides
H. Edison
L. Eeckhoudt
H. Ehrbar
R. Ehrenberg
I. Ehrlich
M. S. Eichenbaum
B. Eichengreen
B. Ellickson
J. Elster
W. Enders
J. Enelow
S. Engerman
D. Epple
T. J. Espenshade
P. D. Evans
C. Evans
D. S. Evans
G. W. Evans
O. Evans
R. E. Evenson
M. L. Eysenbach
J. Fackler
M. Faig
R. C. Fair
H. S. Farber
J. Farrell
G. Feder
D. Feenberg
R. Feenstra
E. Feige
C. Fershtman
G. C. Fethke
A. J. Field
E. B. Field
G. S. Fields
T. W. Fields
S. Figlewski
R. Findlay
F. Fisher
M. Fisher
A. Fishlow
H. Flam
M. J. Flannery
M. A. Flavin
R. P. Flood
N. Folbre
D. K. Foley
T. Follain
J. Formby
R. Forsythe
J. Foster
R. H. Frank
D. Freedman
A. M. Freeman
R. Freeman
J. W. Friedman
K. A. Froot
R. T. Froyen
V. R. Fuchs
J. Fuhrer
D. Fullerton
M. A. Fuss
M. Gaffney
J. E. Gagnon
W. G. Gale
E. Gal-Or
P. M. Garber
B. Gardner
D. Gately
G. Gau
M. Gavin
H. Genberg
T. J. George
M. Gersovitz
M. Gertler
J. Geweke
R. Giammarino
R. S. Gibbons
H. L. Gill
R. G. Gillingham
A. Giovannini
A. Glazer
J. Goddeeris

C. Goetz
V. P. Goldberg
A. S. Goldberger
C. Goldin
M. Goldstein
M. Goodfriend
D. B. Gordon
R. Gordon
R. J. Gordon
W. Gormley
G. Gorton
P. Gottschalk
L. H. Goulder
D. A. Graham
J. W. Graham
J. Gravelle
P. Graves
J. A. Gray
R. C. Green
E. J. Green
R. L. Greenfield
M. L. Greenhut
J. Greenwood
A. Greif
K. B. Grier
J. M. Griffin
Z. Griliches
M. Grinblatt
E. L. Grinols
R. Gronau
E. Groshen
G. M. Grossman
H. I. Grossman
M. Grossman
S. Grossman
H. K. Gruenspecht
A. L. Gustman
J. Gyourko
N. U. Hague
C. Hakkio
R. E. Hall
J. C. Ham
D. S. Hamermesh
B. W. Hamilton
J. D. Hamilton
W. M. Hanemann
T. H. Hannan
L. P. Hansen
W. L. Hansen
E. A. Hanushek
W. S. Haraf
J. E. Harrington, Jr.
G. W. Harrison
O. D. Hart
P. R. Hartley
T. Hatta
J. Haubrich
J. C. Hause
J. A. Hausman
J. Hay
F. Hayashi
J. Heaton
J. J. Heckman
R. Heiner
R. L. Heinkel
P. S. Heller
E. Helpman
P. H. Hendershott
J. V. Henderson
K. Hendricks
W. E. Hendricks
D. Hendry
Z. Hercowitz
B. E. Hermalin
R. J. Herrnstein
J. D. Hess
D. D. Hester
W. Hettich
R. L. Hetzel
D. A. Hibbs
M. Hill
A. L. Hillman
J. R. Hines
J. Hirshleifer
I. J. Hoch
H. M. Hochman
R. J. Hodrick
J. G. Hoehn
J. P. Hoehn
T. Hoerger
S. Hollander
M. Holmer
B. Holmstrom
C. A. Holt
D. Holtz-Eakin
B. E. Honore
P. Hooper
K. D. Hoover
J. L. Horowitz
A. J. Hosios
V. J. Hotz
D. Howard
C. W. Howe
P. Howitt
W. Hoyt
H. J. Hozler
R. G. Hubbard
G. Huberman
C. R. Hulten
T. Humphrey
D. Ingberman
R. P. Inman
R. M. Isaac
Y. Ishii
H. Itoh
C. Jacklin
M. Jacobsen
R. Jacobsen
A. Jacquemin
A. B. Jaffe
E. James
J. A. James
D. W. Jansen
A. John
W. Johnson
B. Jovanovic
P. L. Jowkow
J. M. Joyce
K. L. Judd
J. H. Kagel
V. T. Kaitala
M. I. Kamien
E. C. Kane
E. Karni
J. M. Karpoff
H. Kasper
A. Katz
L. F. Katz
M. Katz
M. L. Katz
J. B. Kau
G. Kaul
H. B. Kazemi
T. E. Keeler
W. R. Keeton
P. Kehoe
T. J. Kehoe
J. F. Kennan
J. Kesselman
I. N. Kessides
N. M. Kiefer
M. Killingsworth
M. S. Kimball
C. P. Kindleberger
M. King
R. G. King
A. Kirman
N. Kiyotaki
M. Klein
P. Kleindorfer
K. M. Kletzer
J. L. Knetsch
M. Knetter
C. R. Knoeber
T. E. Kollintzas
R. C. Kormendi
M. H. Kosters
L. J. Kotlikoff
C. Kowalczyk
J. A. Krautkraemer
I. B. Kravis
D. Kreps
K. Krishna
C. G. Krouse
A. B. Krueger
P. J. Kuhn
R. Kumar-Sah
H. Kunreuther
A. Kuprianov
F. E. Kydland
A. Kyle
A. S. Kyle
H. F. Ladd
J. T. LaFrance
D. Laidler
J. Laitner
R. Lampman
W. Landes
M. Landsberger
K. Lang
L. J. Lau
C. Lave
D. Lavoie
E. Lawrence
S. Layson
E. E. Leamer
J. Ledyard
R. Lee
K. Leffler
A. Leibowitz
L. Leiderman
M. P. Leidy
D. E. Leigh
L. S. Leighton
H. B. Leonard
S. F. LeRoy
D. Levin
D. Levine

J. A. Levinsohn
F. D. Lewis
T. R. Lewis
E. R. Lichtenberg
F. R. Lichtenberg
R. Lind
P. Lindert
B. G. Linster
B. Lipman
S. A. Lippman
R. E. Lipsey
M. B. Loewy
R. E. Lombra
J. B. Long
N. V. Long
R. Lopez
J. R. Lott
M. C. Lovell
R. E. B. Lucas
S. Lundberg
L. J. Maccini
M. Machina
J. K. MacKie-Mason
T. E. MaCurdy
G. S. Maddala
W. Magat
S. Magee
C. Mai
K. Maler
G. Mankin
W. Manning
M. E. Manser
C. Manski
C. Mao
R. A. Margo
N. Marion
S. G. Marks
J. R. Markusen
M. L. Marlow
J. Marquez
D. Marshall
J. M. Marshall
R. E. Martin
A. L. Marty
H. P. Marvel
S. E. Masten
F. Mathewson
K. Matsuyama
S. J. Matusz
W. Mayer
D. Mayers, Jr.
J. Mayshar
R. P. McAfee
J. J. McAndrews
S. McCafferty
B. McCall
J. J. McCall
E. F. McClennen
J. H. McCulloch
J. M. McDowell
M. B. McElroy
D. McFadden
R. A. McGuire
T. J. McGuire
W. McKibbin
J. McMillan
S. K. McNees
E. Meade
S. Megdal
R. Mehra
A. H. Meltzer
R. Mendelsohn
D. L. Mengle
B. D. Meyer
J. Meyer
R. Michener
P. Milgrom
J. H. Miller
M. H. Miller
R. E. Miller
D. Mills
F. Milne
J. Mintz
L. J. Mirman
J. A. Miron
E. J. Mishan
F. S. Mishkin
D. M. Modest
R. Moffitt
J. Mokyr
J. D. Montgomery
P. J. Montiel
D. Mookherjee
D. T. Mortensen
E. Mueller
D. J. Mullineaux
K. J. Murphy
K. M. Murphy
J. C. Musgrave
R. Musgrave
M. P. Mussa
J. Mutti
J. H. Nachbar
B. Nahata
B. Nalebuff
M. P. Narayanan
P. J. Neary
S. Neftci
R. Nelson
R. R. Nelson
M. L. Nerlove
D. Newbery
J. P. Newhouse
S. Nickell
J. Niehans
M. T. Nielsen
S. Nolan
R. G. Noll
W. Nordhaus
W. Oakland
W. Oates
R. L. Oaxaca
A. O'Brien
M. Obstfeld
M. Ogaki
H. Ohta
C. A. Olsen
E. O. Olsen
P. F. Orazem
J. Ordover
M. J. Osborne
J. Ostroy
A. J. Oswald
T. Ozawa
H. Pack
A. Pakes
M. Parkin
B. P. Pashigian
D. Patinkin
M. V. Pauly
C. H. Paxson
S. Peltzman
J. H. Pencavel
M. K. Perry
M. Persson
T. Persson
J. Pesando
P. Pestieau
H. E. Peters
W. Pfahler
C. E. Phelps
L. Phlips
R. S. Pindyck
D. Pines
R. Piron
C. Pitchik
M. W. Plant
C. I. Plosser
R. D. Plotnick
I. P'ng
M. J. Podgursky
T. F. Pogue
K. P. Polenske
A. J. Policano
A. M. Polinsky
J. Pomery
K. T. Poole
W. Poole
R. C. Porter
R. D. Porter
R. H. Porter
P. Portney
J. Poterba
C. Prendergast
G. Prennushi
E. C. Prescott
A. Protopapadakis
F. L. Pryor
D. Purvis
D. Quah
R. E. Quandt
H. C. Quirmbach
R. Ram
V. A. Ramey
A. J. Randall
R. H. Rasche
D. J. Ravenscraft
S. A. Ravid
E. J. Ray
A. Razin
R. Ready
S. Rebolo
W. S. Reece
R. J. Reilly
C. Reimers
J. F. Reinganum
U. E. Reinhardt
P. C. Reiss
A. Reschovsky
G. Rich
J. G. Riley
M. H. Riordan
J. Roback
D. J. Roberts
J. Roberts
R. Roberts
H. Rockoff
D. Rodrik

J. E. Roemer
C. A. Rogers
R. Rogerson
W. Rogerson
K. Rogoff
K. S. Rogoff
R. Roll
C. D. Romer
D. H. Romer
P. M. Romer
T. Romer
S. Rose-Ackerman
H. S. Rosen
K. Rosen
S. Rosen
R. W. Rosenthal
M. R. Rosenzweig
J. Rotemberg
A. E. Roth
N. Roubini
D. L. Rubinfeld
G. D. Rudebusch
J. R. Rudin
R. J. Ruffin
D. Runkle
M. Rush
T. Russell
J. Rust
J. Sachs
Z. Safra
B. Sahni
G. Salamon
S. Salant
M. K. Salemi
M. A. Salinger
G. Saloner
L. Samuelson
W. F. Samuelson
S. H. Sandell
T. Sandler
A. M. Santomero
D. E. M. Sappington
T. J. Sargent
R. Sato
M. A. Satterthwaite
A. Saunders
D. Scheffman
F. M. Scherer
D. E. Schlagenhauf
R. L. Schmalensee
P. J. Schmidt
R. Schmidt
J. A. Schmitz, Jr.
K. Scholz
M. Schulenburg
T. P. Schultz
W. D. Schulze
P. M. Schwarz
U. P. Schweizer
G. W. Schwert
S. Scotchmer
L. Scott
G. W. Scully
J. J. Seater
U. Segal
G. Shaffer
A. Shaikh
A. Shaked
C. Shapiro
D. Shapiro
M. D. Shapiro
P. Shapiro
S. Shavell
K. Shaw
S. M. Sheffrin
A. Shepard
R. Sherman
J. Shih
R. J. Shiller
Y. Shinkai
A. Shleifer
J. B. Shoven
W. F. Shughart II
O. Shy
A. Sibert
D. Sibley
D. R. Siegel
E. Silberberg
C. Simon
C. A. Sims
K. J. Singleton
H. W. Sinn
A. Siow
J. Skinner
S. Sklivas
M. E. Slade
J. Slemrod
A. D. Slivinski
F. Sloan
S. M. Slutsky
D. H. Small
K. A. Small
A. Smith
B. Smith
G. Smith
J. P. Smith
R. S. Smith
S. Smith
V. K. Smith
V. L. Smith
A. Snow
D. J. Snower
J. M. Snyder
R. A. Snyder
J. Sobel
G. R. Solon
J. Sonstelie
R. H. Spady
C. Spatt
J. Spaulding
S. Spear
B. J. Spencer
M. W. Spicer
D. F. Spulber
P. Srinagesh
G. Stadler
F. P. Stafford
R. Staiger
D. A. Starrett
R. Startz
J. C. Stein
L. Stekler
N. H. Stern
G. J. Stigler
J. H. Stock
G. L. Stockhausen
A. C. Stockman
D. J. Stockton
T. M. Stoker
N. L. Stokey
R. Strauss
C. Stuart
R. M. Stulz
D. Suits
D. G. Sullivan
M. A. Sullivan
L. Summers
J. Sutton
L. E. O. Svensson
V. Swaroop
J. Swierzbinski
A. O. Sykes
G. Tabellini
E. Tanner
B. Taub
P. Taubman
H. Tauchen
J. B. Taylor
L. Taylor
T. Teitenberg
R. Thaler
S. E. Thiel
J. F. Thisse
D. Thomas
J. P. Thomas
L. G. Thomas
W. Thomson
N. Tideman
M. Todaro
R. H. Topel
E. Tower
S. J. Trejo
R. Tresch
J. H. Tressler
R. Triest
R. W. Tryon
G. Tullock
S. J. Turnbull
S. J. Turnovsky
A. Turunen-Red
A. Tversky
D. Usher
J. C. Van Huyck
E. C. H. Veendorp
C. Vellturo
S. F. Venti
W. K. Viscusi
R. W. Vishny
G. von Furstenberg
P. A. Wachtel
K. W. Wachter
A. Wagstaff
D. G. Waldo
J. R. Walker
H. Wan
M. Waterson
M. W. Watson
R. H. Webb
L. Wegge
K. C. J. Wei
D. A. Weil
P. Weil
E. R. Weintraub
B. Weisbrod
A. Weiss
L. Weiss
T. Weisskopf
M. L. Weitzman

F. R. Welch
W. P. Welch
S. H. Wellisz
J. T. Wenders
K. West
J. A. Weymark
J. Whalley
W. C. Wheaton
M. D. Whinston
J. K. Whitaker
L. H. White
M. J. White
C. Whiteman
S. N. Wiggins
R. Wigle
D. Wilcox
D. Wildasin
L. L. Wilde
T. D. Willett
J. T. Williams
O. Williamson
S. D. Williamson
J. D. Wilson
R. B. Wilson
S. Winer
R. A. Winter
R. Wintrobe
M. Wiseman
A. D. Witte
B. L. Wolfe
J. R. Wolfe
E. Wolff
A. Wolinsky
K. Wolpin
S. A. Woodbury
M. Wooders
A. D. Woodland
M. Wynne
L. B. Yeager
J. Yinger
S. Yitzhaki
L. Young
E. E. Zajac
A. Zardkoohi
G. A. Zarkin
V. Zarnowitz
J. Zax
R. Zeckhauser
S. P. Zeldes
D. Zilberman
A. Zimbalist
S. Zin
J. Zinser

ORLEY ASHENFELTER, *Editor*

Report of the Editor

Journal of Economic Literature

The mission of the *Journal of Economic Literature* is to help members of the Association keep abreast of research developments in various fields of economics. This goal is effected by providing a bibliographic guide to research publications, reviews of certain books, and articles describing and evaluating research progress on particular topics.

The *Journal*'s bibliographic departments (i.e., the Annotated Listing of New Books, the Contents of Current Periodicals, Selected Abstracts, and the list of Doctoral Dissertations in Economics) are run from our Pittsburgh office under the direction of Drucilla Ekwurzel. Mary Kay Akerman serves as Assistant Editor and Asatoshi Maeshiro is an Editorial Consultant with general responsibility for the classification of articles in the *Journal*'s subject index. The Pittsburgh office benefits also from the services of Patricia Andrews, Elise Braden, Elizabeth Braunstein, and Elizabeth Thornton. They have my thanks for their valuable efforts throughout the year.

The Pittsburgh office is responsible also for the *Index of Economic Articles* and the Economic Literature Index available on Dialog Information Retrieval Services. In 1989, the 1985 *Index* was published and shipped, and the 1986 volume is due to be shipped at the end of 1989. The 1987 *Index* should be published in mid-1990 with the 1988 *Index* following about six months after. The *Journal* now handles the marketing of the *Index* and also receives royalties from the use of the Economic Literature Index. In this way, our bibliographic work is generating substantial revenues for the Association. These may be augmented in the future if we are successful in developing and merchandising a CD-ROM version of our bibliographic data. With the bibliographic information becoming increasingly available in forms other than the quarterly issues of the *Journal*, it may be appropriate in the future to economize on our production expenses and reduce the volume of such material published in the *Journal*. We routinely seek cheaper ways of delivering services to our members. In fact, as shown in Table 1, the total size (number of pages printed) of the *Journal* has not grown over five years.

During the past year, we have been reviewing and revising the Classification System for Articles and Abstracts. As mentioned in last year's Report, there is a

TABLE 1—*JEL* PAGES BY DEPARTMENT, 1981–89

	Articles & Communications	Book Reviews	New Book Annotations	Current Periodicals[a]	General Index	Total
1980	366	294	276	1,072	26	2,034
1981	342	286	270	1,059	23	1,980
1982	331	251	300	1,069	23	1,974
1983	305	239	281	1,086	38	1,949
1984	354	225	314	1,193	37	2,124
1985	364	237	299	1,306	38	2,244
1986	367	250	308	1,344	38	2,307
1987	345	251	315	1,352	40	2,303
1988	419	241	318	1,240	40	2,258
1989	334	251	328	1,254	41	2,208

[a] In 1987, the *Journal of Economic Literature* took over from the *American Economic Review* the responsibility of publishing the list of Doctoral Dissertations in Economics. This item is added to "Current Periodicals" which also includes the Contents of Current Periodicals, the Subject Index of Articles in Current Periodicals, and Selected Abstracts.

TABLE 2—PROPOSED GENERAL CATEGORIES[a]

A	General Economics (010, 0200)
B	Methodology and History of Economic Thought (030)
C	Mathematical and Quantitative Methods (200)
D	Microeconomics (021, 022, 024, 025, 026, 920)
E	Macroeconomics (023, 130)
F	International Economics (400)
G	Financial Economics (310)
H	Public Economics (320)
I	Health, Education, and Welfare (910)
J	Labor and Demographic Economics (800)
K	Law and Economics (916)
L	Industrial Organization (610, 630, 640)
M	Business Economics and Accounting (500)
N	Economic History (040)
O	Economic Development, Technological Change, and Growth, (110, 121, 122, 620)
P	Economic Systems (027, 050, 113, 123, 124)
Q	Agricultural and Natural Resource Economics (700)
R	Urban, Rural, and Regional Economics (930, 940)
Z	Miscellaneous

[a] The numbers in parentheses relate to the existing Classification System.

widespread feeling that, in certain areas, the current System does not accommodate current economics research and publications well and, therefore, we have canvassed a wide spectrum of opinion to devise a better System. We have drawn up nineteen general categories as shown in Table 2, and we are in the process of specifying subcategories within these larger groupings. We hope to introduce and operate the new Classification System next year.

The Articles and Communications and Book Review departments of the *Journal* are managed from the Department of Economics at Stanford University. During 1989, the *Journal* published 9 major articles, 1 communication, and 164 book reviews. Alex Field supervises the Book Review department and Moses Abramovitz shares with me the editorial duties associated with reviewing the manuscripts. We receive excellent support from Anita Makler, Ann Vollmer, and Toni Haskell. They all have my thanks for their significant contribution to the *Journal*'s success.

In addition, we have benefited from generous advice given by our Board of Editors. Three members of the Board have completed their terms as of the end of 1989, namely, Duncan Foley, Mark Killingsworth, and Roger Noll. I am most grateful for their work. Roger Noll deserves special mention not only for having served since 1981, but also for remarkably unstinting and constructive help. I am very pleased that Duncan Foley has agreed to serve another term on the Board. I shall be proposing to the AEA Executive Committee that Charles Brown of Michigan University and John Riley of UCLA join the Board next year. I thank all the members of the Board for the energy and time they have devoted to the *Journal*. Also we have called upon a number of referees, many of whom have made very substantial contributions to the quality of our articles. Their efforts are much appreciated. We list below the referees used during 1988 and 1989.

A. B. Abel
G. A. Akerlof
B. Allen
F. Allen
T. W. Anderson
E. S. Andrews
K. Arrow
B. W. Arthur
O. Attanasio
Y. Barzel
B. Bateman
B. Bernanke
R. A. Berry
N. Birdsall
D. Bizer
F. D. Blau
T. E. Borcherding
G. Borjas
M. J. Boskin
D. F. Bradford
G. F. Break
J. Bulow
J. Y. Campbell
D. Card
C. Carson
V. Chick
R. S. Chirinko
G. C. Chow
R. L. Clark
R. Cornwall
B. Craig
S. B. Danziger
P. A. David
D. G. Davies
L. DeAlessi
N. B. DeMarchi
R. A. Easterlin
R. Eisner
R. J. Epstein
P. Evans
J. Evensky
R. C. Fair
S. M. Fazzari
G. S. Fields
A. C. Fisher
V. Fuchs
R. E. Gallman
M. Gersovitz
R. J. Gilbert
A. S. Goldberger
R. J. Gordon
J. Gurley
R. Halvorsen
B. W. Hamilton
R. H. Haveman
J. V. Henderson

C. Hildreth
F. M. Howland
G. Jaynes
G. Johnson
B. F. Johnston
P. Joskow
T. E. Josling
F. T. Juster
J. Kalt
L. Katz
J. W. Kendrick
N. M. Kiefer
S. R. King
J. Kornai
A. O. Krueger
P. Krugman
P. W. Kuznets
F. E. Kydland
N. Lardy
L. J. Lau
E. E. Leamer
R. D. Lee
H. M. Levin
S. Lewis
S. J. Liebowitz
J. Litwak
M. Machina
A. Manne
J. W. Marlin, Jr.
F. Matthewson
R. C. O. Matthews
R. I. McKinnon
C. E. McLure, Jr.
J. Mellor
R. T. Michael
O. S. Mitchell
R. Moffitt
A. Munnell
R. R. Nelson
E. Neuberger
P. K. Newman
W. D. Nordhaus
J. R. Norsworthy
W. Novshek
J. D. Owen
J. P. Palmer
D. O. Parsons
H. T. Patrick
K. Pavitt
J. K. Peskin
A. M. Polinsky
G. Psacharopoulos
G. Ranis
T. G. Rawski
P. K. Robins
K. Rogoff
V. V. Roley
T. Romer
W. W. Rostow
A. Roth
J. A. Roumasset
N. Ruggles
R. Ruggles
J. D. Sachs
T. Sargent
I. V. Sawhill
J. Scadding
A. Sen
M. D. Shapiro
R. J. Shiller
E. Silberberg
C. A. Sims
J. Skinner
T. M. Smeeding
J. P. Smith
R. Solow
F. P. Stafford
R. Staiger
O. Stark
W. E. Steinmueller
D. A. Sumner
R. C. Sutch
P. Timmer
J. Tobin
J. Vanek
B. Ward
M. Warshawsky
M. Watts
M. L. Weitzman
H. White
D. Whittman
J. G. Williamson
O. E. Williamson
F. A. Wolak, Jr.
E. N. Wolff
G. Wright

JOHN PENCAVEL, *Editor*

Report of the Editor

Journal of Economic Perspectives

By the end of 1989, the *Journal of Economic Perspectives* will have broken into double figures by publishing ten issues. Judging from the formal and informal responses the Editors have received, it seems to have carved out a niche for itself in the world of economics journals.

The main formal feedback the *Journal* has received was from an *Ad Hoc* Committee of the AEA, headed by William Baumol, that was set up to examine the contents and financing of this journal. To help inform its deliberations, the Committee carried out a survey of about 1000 AEA members in mid-1988. The final report of the Committee was presented to the Executive Committee in March 1989.

Among other findings, the Committee noted: "Most members of the AEA like the *JEP*. Its average approval rating is not significantly different from that accorded to the *AER* or *JEL*." This finding was especially encouraging to the Editors since, at the time of the survey, respondents had seen only three or four issues of this journal. The Baumol report recommended that "the Executive Committee of the AEA reaffirm its support of the *JEP*," and the Executive Committee did so at its March 1989 meeting.

In the four issues of 1989, the *Journal* has continued to touch upon a broad range of topics in a broad range of formats. The issues of 1989 included seven symposia: women in the labor market, the budget deficit, auctions, real business cycles, the savings and loan crisis, seniority and wages, and social norms. They also included 5 articles about famous economists (Sen, two on Solow, Schelling, and Hicks); 15 articles on various other topics, ranging from chaos theory and experimental economics to corporate bankruptcy and environmental economics to the role of potential competition and the political economy of trade policy, and others. In addition, the *Journal* also continued publication of three regular features ("Puzzles," "Anomalies," and "Recommendations for Further Reading") along with Notes and Correspondence. In 1990, the *Journal* intends to start two additional features, one of which will appear in each issue: "Retrospectives" on topics in the history of economic thought, and "Policy Watch" on topics of current policy interest.

In total, the 1989 issues of the *Journal* included 832 pages, consisting of 38 articles, 5 introductions or very brief articles, 11 of the features articles, plus Notes and Correspondence.

As this rich and varied flow of material indicates, the decentralized editorial procedures of the *Journal* have continued to work well. The Associate Editors have produced a steady stream of ideas and authors, enabling the *Journal* to draw on a large group of economists from different geographical areas and with diverse interests and perspectives. After several years of operation, a process of natural turnover among the Associate Editors seems to have begun, thus allowing several new members to bring their ideas to the *Journal* each year.

In fact, the Associate Editors have been so diligent that the issues of 1990 are almost fully booked already, and advance planning for 1991 is well underway. The Editors are delighted to report that this high level of output has not meant any decrease in quality. Rather the reverse seems to be true: after ten issues, authors for the *Journal* now have models of the sort of exposition the *Journal* is seeking, and they are working hard to live up to the standards of previous authors.

The incumbent Associate Editors who will continue into 1990 are: Henry Aaron, The Brookings Institution; Pranab Bardhan, University of California-Berkeley; Dwight Jaffee, Princeton University; Edward Lazear, University of Chicago and The Hoover Institution; Mark Machina, University of California-San Diego; N. Gregory Mankiw, Harvard University; John Roemer, University of California-Davis; Kenneth Rogoff, University of California-Berkeley; Bernard Saffran, Swarthmore College; Steven Salop, Georgetown University; Lawrence Summers, Harvard University; Hal Varian, University of Michigan; Gavin Wright, Stanford University; Janet Yellen, University of California-Berkeley.

Members of the *Journal*'s Advisory Board have also continued to provide helpful advice on matters of editorial policy, as well as by offering feedback on past articles and occasional suggestions for future ones. Each year, the December meetings have provided an opportunity for members of the Advisory Board to meet with the Associate Editors, as well as the Editor, Co-Editor, and Managing Editor, to exchange views concerning the journal.

During 1989, the position of Editorial Associate at the *Journal* was performed by Debbie Sachs. She managed to remind our authors of their deadlines; to keep the *Journal* up to date with our typesetter and printer; to deal with much of the Stanford University bureaucracy as needed; to edit the Notes; and to do all the other small details that keep the office running on a day-to-day basis. Her performance and attention to detail have been exemplary.

The Editors again feel that they cannot overstate the important role that Managing Editor Timothy Taylor has played in the operation of the *Journal*. He has continued to ensure that the fundamental objectives of the *Journal* are satisfied. He has performed the difficult task of persuading authors to amend and rewrite their articles with delicacy and skill, and has shown that it is possible to edit papers and increase their clarity and accessibility, while still retaining the distinctive voice of each author.

Requested actions: 1) Approval of new members to join the Editorial Board in 1990: James Heckman, Yale University; Barry Nalebuff, Yale University; Sherwin Rosen, University of Chicago.

JOSEPH STIGLITZ, *Editor*
CARL SHAPIRO, *Co-editor*

Report of the Director

Job Openings for Economists

The total number of new jobs listed increased to 2,014, a new record, exceeding by 47 the old record of 1,967 set last year. Academic jobs increased from 1,387 to 1,411, and nonacademic listings increased from 580 to 603. Table 1 shows total listings (employers), total jobs, new listings and new jobs by type (academic and nonacademic) for each issue of *JOE* in 1989.

Table 2 shows the number of employers by category (four-year colleges, universities with graduate programs, federal government,

TABLE 1—JOB LISTINGS FOR 1989

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	82	149	71	127
April	44	84	38	71
June	26	44	24	38
August	49	102	45	98
October	200	548	177	511
November	133	302	133	302
December	196	464	96	264
Subtotal	730	1,693	584	1,411
Nonacademic				
February	16	66	11	48
April	23	84	19	69
June	17	47	14	36
August	29	93	25	77
October	53	205	45	174
November	26	71	26	71
December	58	242	29	128
Subtotal	222	808	169	603
Total	952	2,501	753	2,014

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1989

Issue	Four-year Colleges	Universities with Graduate Programs	Federal Govern-ment	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	29	43	2	1	3	1	8	1	98
April	14	30	2	5	6	2	6	2	67
June	6	20	4	5	2	–	15	1	43
August	13	36	6	3	7	1	10	2	78
October	73	127	14	4	10	1	17	7	253
November	51	82	10	–	5	2	9	–	159
December	82	114	17	1	10	4	22	4	254
Totals	278	458	55	19	43	11	77	17	952

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1989

Fields	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	71	44	22	58	270	137	246	848
Growth and Development (100)	26	16	9	23	55	28	52	209
Econometrics and Statistics (200)	30	23	18	26	101	47	91	336
Monetary and Fiscal (300)	32	23	13	31	118	57	111	385
International Economics (400)	26	19	8	28	94	52	95	322
Business Administration, Finance, Marketing and Accounting (500)	35	28	6	15	73	33	80	270
Industrial Organization (600)	23	23	7	33	88	32	87	293
Agriculture and Natural Resources (700)	13	13	17	24	38	11	32	148
Labor (800)	22	14	8	13	53	32	53	195
Welfare and Urban (900)	20	17	3	22	57	44	57	220
Related Disciplines (A00)	1	–	–	1	11	9	7	29
Administrative Positions (B00)	9	6	2	4	12	11	16	60
Totals	308	214	113	278	970	550	936	3,315

Note: Fields of specialization codes are from the *Journal of Economic Literature*.

etc.) for each of the seven issues. Academic institutions continue to be the major advertisers of jobs, about 75 percent of the total number of employers listing vacancies. The pattern of employers by type is roughly the same as it has been for the last several years.

Table 3 shows the number of citations by field of specialization. Once again, General Economic Theory (000) led, Monetary and Fiscal (300) was second, and Econometrics and Statistics (200) third; this pattern has existed for many years.

Violet Sikes handles everything to do with the publication and distribution of *JOE*. I am very grateful for her excellent work. She is due full credit for the success of the publication.

C. ELTON HINSHAW, *Director*

Report of the Committee on Economic Education

The American Economic Association created a Committee on Economic Education (CEE) in the mid-1950s, which grew to maturity by the late 1960s. The beginning occurred at the Spring 1952 AEA Executive Committee meeting, when it was voted to accept membership on the Commission on Economics in [Elementary and Secondary Level] Teacher Education, and in light of this experience to investigate the desirability of further cooperation with the Joint Council on Economic Education (JCEE). Later that year, the Executive Committee voted to establish an official *Ad Hoc* Committee on Economics in Teacher Education. This committee was charged with concern about precollege economics education.

At the December 1955 meeting of the Executive Committee, the *Ad Hoc* Committee on Teacher Education was recast as a standing Committee on Economic Education. The purpose of this standing committee was to serve as a focal point for improving the status of economic education within the field of professional economists, to stimulate and encourage professional work on economic education, and to encourage the preparation of articles about and arrange sessions at the annual AEA meetings on economics education. But, even from the beginning, "there was debate concerning [the Committee's] functions." (Minutes of the Executive Committee Meeting of December 28, 1955, *American Economic Review Proceedings*, May 1956, p. 589.)

Sessions on the teaching of economics in the schools were included on the 1955 and 1956 AEA annual meeting programs and published in the succeeding years' *AER Proceedings*. By special request of the CEE, sessions on economics education occurred again at the 1960 and 1961 annual meetings. During its early years, the CEE focused on precollege teacher training and developed a formal affiliation with the JCEE. Other activities of the period included preparation of a roster of economists interested in economics education, and a study of textbooks used in secondary school courses in American history, social problems, and economics. In 1960, the Committee helped establish the National Task Force on Economic Education whose goal was to describe the minimum level of economic understanding essential for the practice of good citizenship. The Task Force's report, *Economic Education in the Schools*, described what economics should be taught at the high school level. The committee also assisted in producing *The American Economy* television series aired nationwide by CBS and PBS in 1963.

After a careful review of the role of the AEA in economics education, the Executive Committee in December 1963 dissolved the Committee on Economic Education and created a successor with the same name. This committee was to be more action oriented. It was charged with improving the quality of economics education at all levels, from precollege to college, adult, and general economics education. By request of the CEE, the Executive Committee of the AEA in March 1964 voted to include a special session on economics education at each annual AEA meeting. With the exception of 1965, there has been a special session on economics education at each annual meeting of the AEA since 1964.

In 1961 the National Task Force on Economic Education recommended that the JCEE construct a standardized Test of Economic Understanding. A specially appointed Committee on Measurement of Economic Understanding completed this high school level exam (renamed the Test of Economic Literacy in its revised forms) in 1964. Another blue ribbon committee of distinguished economists was then asked to assemble a similar exam for the introductory college level course. The college level exam, called the Test of Understanding College Economics (TUCE) was completed in 1968 and has since been revised in 1979 and 1989. The TUCE fostered rapid growth in research on teaching college economics in the 1970s, as it provided a standardized measure of "output." The CEE also helped the JCEE establish the *Journal of Economic Education*, which published its first issue in 1969, and has been disseminating important research findings and information about teaching economics ever since.

In the 1970s, the CEE turned its attention to teacher-training programs, establishing in the late 1970s national workshops on economics instruction and helping the JCEE produce material to be used with such workshops (*Resource Manual for Teacher Training Programs in Economics*. JCEE, 1978). The CEE also encouraged the *Journal of Economic Literature* to publish a survey of research findings on the teaching of economics, and it cosponsored a survey of the economics curriculum in U.S. four-year colleges and universities.

Finally, in the 1980s, the CEE has been active in developing an advanced placement examination in economics, nurturing young scholars interested in doing research on teaching high school economics, sponsoring a conference on textbooks for the principles course, and continuing work on the economics major. All of these efforts have been cooperative with the JCEE.

In August 1989, the CEE convened in Washington D.C. for two days to develop an agenda for its activities in the 1990s. At this meeting, it was agreed that "promoting effective economics education" at both the precollege and college level is an undisputed objective of the CEE. Committee members argued that the demand for publicly provided services like those produced by the committee should be grounded in some private market failure. The Committee has historically emphasized the creation and dissemination of information about effective teaching methods. This fits the market failure criterion, as information is frequently subject to scale economies, free-rider and externality problems, and is available to additional users at close to zero marginal cost.

With these criteria in mind, the Committee outlined a new agenda. Details about this agenda will be reported at next year's annual AEA meetings session on economics education. Here are listed six broad responsibilities that have been and will continue to be discharged by the AEA Committee on Economic Education:

1) Enhancing recognition within the profession of the importance of effective teaching; 2) Facilitating the training of more effective economics teachers; 3) Improving the assessment of what and how much our students learn; 4) Cooperating with and assisting the JCEE (with its primary mission of improving precollege economics education); 5) Encouraging innovation in teaching methods and research on teaching effectiveness; 6) Disseminating information about effective economics teaching.

JOHN J. SIEGFRIED, *Chair*

Report of the Committee on the Status of Minority Groups in the Economics Profession

This report summarizes the major activities under the perview of the Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) for 1989. These activities cover the AEA Summer Minority Program, the AEA/FRS minority graduate fellowship program, and fundraising activities.

AEA Summer Minority Program. In 1990, the AEA Summer Minority Program will be offered for the seventeenth consecutive year. This program was established in order to attract members of underrepresented, disadvantaged minority groups (Native Americans, blacks, and Hispanics) to graduate study in economics. The AEA Summer Program will be hosted by Temple University in 1990 for the fifth year. Michael Leeds is the current director of the summer program.

The CSMGEP is now in the process of selecting the university it will recommend to the AEA Executive Committee as the next host for the summer program. Selection will be based on proposals submitted to the AEA by academic institutions interested in hosting the program.

In cooperation with Temple University, the CSMGEP is compiling an updated list of past AEA Summer Fellows. A questionnaire has been sent to individuals on the current mailing list seeking such information as current activities and addresses. A list of past participants for whom no current mailing address is available was included in the survey in an attempt to locate them through a former classmate. They revised mailing list will be used to distribute periodic questionnaires that will enable CSMGEP to conduct an ongoing evaluation of the summer program, and will serve as the basis for a newsletter for summer program alumni.

AEA/FRS. The AEA/FRS program is directed toward supporting minority graduate students for a two-year period after they have successfully completed their Ph.D. preliminary exams and have begun the dissertation. The program is intended to assist minority graduate students who are writing dissertations in areas of interest to the Federal Reserve. New fellowships for the 1989–90 academic year were awarded to Sandra D. Cooke at Howard University and Victor Mendez at Syracuse University. Continuing support was approved for Kenneth Daniels at the University of Connecticut-Storrs, Lorraine Howard at the University of Pennsylvania, Philip Jefferson at the University of Virginia, and Ana Zapata at Fordham University. Rhonda Grass at Howard University and Carlos Elias at New York University completed their Ph.D. degrees in 1989.

Fund Raising. Current support for the summer program is derived mainly from a grant by the Ford Foundation supplemented by small grants from Federal Reserve Banks. The AEA/FRS graduate fellowship program is entirely funded by the Federal Reserve System. The CSMGEP is actively seeking funding to establish a graduate fellowship program to support minority graduate students in their first two years of graduate study in economics. This funding is needed to replace the expired AEA/Rockefeller program, and to bridge the gap between the summer program and the AEA/FRS minority graduate fellowship program. Ongoing fund raising efforts are also being conducted on behalf of the summer program to ensure its financial stability.

RONALD L. OAXACA, *Chair*

Report of the Committee on the Status of Women in the Economics Profession

The American Economic Association's charge to the Committee on the Status of Women in the Economics Profession (CSWEP) is to monitor the position of women in the profession, and undertake activities to improve it. This Report briefly examines the position of women economists in academia from 1974 to 1988 and describes the Committee's recent activities.

The Changing Status of Women Economists in Academia

Since 1974, there has been a strong upward trend in the proportion of faculty in graduate economics departments who are women, although the proportion of those receiving tenure remains somewhat lower than might be expected based on these trends. These results are based on data about graduate economics departments (i.e., departments that grant Ph.D.s) that responded to the AEA's Universal Academic Questionnaire (UAQ) in any year between 1974 and 1988.[1] Because few departments responded to the questionnaire in every year, the data used in this analysis for each year are based on all graduate departments that responded in that year.[2] Although a large matched sample of departments would be preferable for this longitudinal analysis, the relative measures used here (such as the proportion of assistant professors who are women) would not be greatly affected by the year-to-year variation in the group that did respond as long as this variation was not systematically related to the presence of female faculty. If, however, the departments that responded later in the period tended to be the ones that had hired and promoted the most women, these data would overstate the progress being made by female economists.[3]

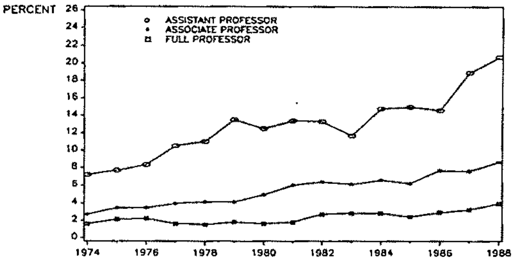


FIGURE 1. PROPORTION OF FACULTY IN GRADUATE DEPARTMENTS WHO ARE WOMEN, BY RANK: 1974–88
(As a percent of faculty with the same rank)

As shown in Figure 1, the proportion of assistant professors who were women tripled between 1974 and 1988—rising from 7 to 21 percent. This increase reflected the growth in the proportion of Ph.D.s in economics awarded to women from under 10 percent in the early 1970s to about 20 percent near the end of the 1980s. The proportion of associate professors who were women also tripled, from 3 to 9 percent, while the proportion of full professors more than doubled from under 2 to almost 4 percent. In both cases, this growth reflected the rising proportion of women at the next lower rank.

Similar patterns are apparent in looking at promotions to associate and full professor. Figure 2 shows the proportion of faculty promoted to associate professor, and the proportion promoted to full professor, who were women. Of those becoming associate

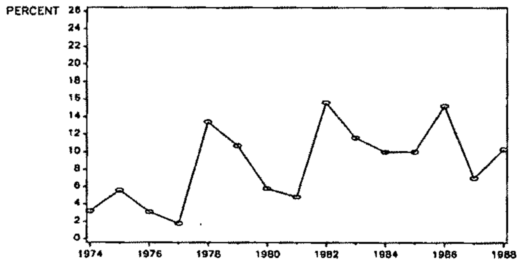
---

[1] The Committee thanks Charles Scott and Joan Haworth for their valuable contributions to this analysis. Please note that, as described later in the text, the UAQ is the source for the figures and tables herein.

[2] The number of departments responding declined over the 1974–88 period, however, from an average of 90 during the 1970s to an average of 70 in the past 6 years.

[3] The longitudinal results presented here are consistent with information based on two particular matched samples. One sample, consisting of departments that responded in both the 1980–81 and the 1987–88 school years, was examined in last year's annual report (see *AER Proceedings*, May 1989, pp. 422–25). Data about departments that responded in both the 1984–85 and the 1988–89 school years are analyzed at the end of this section.

PROMOTION FROM ASSISTANT TO ASSOCIATE PROFESSOR



PROMOTION FROM ASSOCIATE TO FULL PROFESSOR

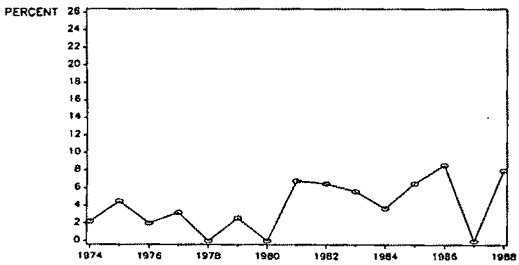


FIGURE 2. PROPORTION OF FACULTY PROMOTED TO THE NEXT RANK WHO ARE WOMEN, GRADUATE DEPARTMENTS ONLY, 1974–88

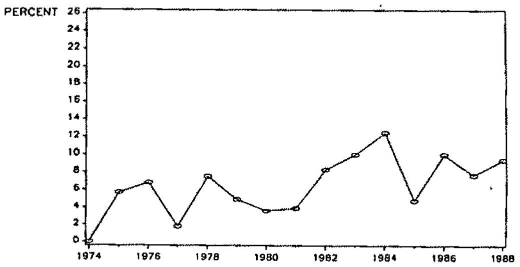


FIGURE 3. PROPORTION OF FACULTY RECEIVING TENURE WHO ARE WOMEN, GRADUATE DEPARTMENTS ONLY: 1974–88

professors between 1974 and 1978, 6 percent, on average, were women. This proportion rose to about 9 percent in the 1979–83 period, and to 10 percent in the 1984–88 period. Of those becoming full professors in the three 5-year intervals, an average of 2, 4, and 5 percent, respectively, were women.

The proportion of faculty receiving tenure who were women is shown in Figure 3. Again, the progress of women in the economics profession is apparent: this proportion averaged 4, 6, and 9 percent in the three 5-year periods. Nonetheless, these figures are somewhat lower than might be expected, considering that the proportion of assistant professors who were women has been over 10 percent since 1977 and at least 13 percent since 1979.

Progress over the past 4 years may also be examined using data from the departments responding to the UAQ in both the 1984–85 and the 1988–89 school years, although the results must be treated with caution because the two academic years may not be representative ones.[4] As shown in Table 1, women received a higher proportion of the Ph.D.s awarded in 1988 than in 1984, but represented about the same proportion of graduate students in both years. Female Ph.D. students were about as likely as their male counterparts to receive financial aid, although female Master's students were more likely than their male counterparts to be assisted in both years.

Women made up a higher proportion of the faculty in 1988 than in 1984 in both undergraduate and graduate economics departments in the matched sample. Table 2 shows that growth occurred at all ranks except that of full professor. In fact, because there were more full professors in these departments in 1988 than in 1984 (718 compared with 672, for a net gain of 46), the number of women at that rank actually rose from 35 to 36 even though their proportion declined slightly. Nonetheless, this increase of one woman stands in sharp contrast to the increase of 45 men with the rank of full professor in these departments.

Table 3 provides information about the hiring, promotion, and tenure decisions of

---

[4] The matched sample contains 78 of the 159 undergraduate institutions that responded in 1988, 19 of the 38 institutions that award only Bachelor's and Master's degrees, and 45 of the 66 institutions that also award Ph.D.s. Although the matched sample includes only one-half of all institutions reporting in 1988, the characteristics for the entire group are remarkably similar to those for the sample analyzed here. The figures presented here are available on request for the entire 1988 sample.

TABLE 1—SELECTED DATA ON STUDENTS IN ECONOMICS

	1984-85			1988-89		
	Total No.	No. Female	Percent Female	Total No.	No. Female	Percent Female
Recipients of Degrees in Economics[a]						
Bachelor's	6,595	2,150	32.6	7,546	2,562	34.0
Master's	536	151	28.2	566	153	27.0
Ph.D.	304	46	15.1	381	80	21.0
Graduate Students in Economics[b]						
Enrolled						
In Ph.D. Program	2,338	499	21.3	2,341	501	21.4
In MA Program	701	177	25.2	552	143	25.9
Receiving Aid						
In Ph.D. Program	1,648	358	21.7	1,784	386	21.6
In MA Program	229	78	34.1	177	55	31.1
Distribution of Employment for Ph.D. Recipients (in percent)						
	Male		Female	Male		Female
Education	44		56	49		53
Government	8		11	8		20
Private Sector	24		19	10		8
Other	24		15	34		18
Total	100		100	100		100

[a]From 78 institutions awarding only Bachelor's degrees, 19 awarding Bachelor's and Master's degrees, and 45 that also award Ph.D.s.

[b]Full-time students attending the 45 institutions in the matched sample that award Ph.D.s.

TABLE 2—NUMBER OF FULL-TIME FACULTY AND PROPORTION WHO ARE WOMEN, BY TYPE OF INSTITUTION AND RANK OF FACULTY[a]

	1984-85			1988-89		
	Total No.	No. Female	Percent Female	Total No.	No. Female	Percent Female
Undergraduate Institutions						
Full Prof.	105	15	14.3	130	16	12.3
Associate Prof.	107	12	11.2	129	17	13.2
Assistant Prof.	152	24	15.8	143	29	20.3
Instructor	47	8	17.0	24	8	33.3
Total	411	59	14.4	426	70	16.4
Graduate Institutions						
Full Prof.	567	20	3.5	588	20	3.4
Associate Prof.	270	15	5.6	230	21	9.1
Assistant Prof.	320	48	15.0	313	64	20.4
Instructor	28	3	10.7	18	2	11.1
Total	1,185	86	7.3	1,149	107	9.3

[a]Excludes 19 institutions that award Bachelor's and Master degrees, but not Ph.D.s.

TABLE 3—CHANGES IN STATUS OF FULL-TIME FACULTY AT PH.D.-GRANTING GRADUATE INSTITUTIONS

	1984-85			1988-89		
	Total No.	No. Female	Percent Female	Total No.	No. Female	Percent Female
Hired						
As Assistant Prof.	54	13	24.1	67	9	13.4
As Associate or Full Prof.	22	1	4.5	23	3	13.0
Promoted						
To Associate Prof.	23	2	8.7	23	2	8.7
To Full Prof.	19	1	5.3	18	0	0.0
Tenured						
At Associate Prof.	20	3	15.0	18	1	5.6
At Full Prof.	4	0	0.0	2	0	0.0

the graduate departments in the matched sample. Women made up a smaller proportion of newly hired assistant professors, but a larger proportion of newly hired associate or full professors, in 1988 than in 1984. Although it may appear that there were substantial differences regarding promotion and tenure between the two years, small numbers of events are the actual explanation. For example, the 5.3 percent of promotions to full professor in 1984 represented one woman, compared with none in 1988.

In summary, as female economists progress through their careers, they are advancing in rank and becoming tenured in greater numbers. A basic question remains, however. How fast should the proportion of associate and full professors who are women grow? The answer depends on the total number of positions at these levels, the rate of turnover caused by factors such as retirement, and the quality of particular male and female faculty members being considered for promotion. The CSWEP Board hopes to be able to shed some light on these questions in next year's Report.

The Committee's Recent Activities

The Committee's activities in 1989 continued to focus on helping women advance in the economics profession. To expand the number of entries in CSWEP's roster of female economists, chairs of economics departments were asked to provide questionnaires and membership applications to their female graduate students and faculty, as well as to any men they thought might want to join CSWEP.[5] The chairs were also asked to send the same information to any recent graduates whose current addresses were available. The Board is extremely grateful for their cooperation which has resulted in large numbers of new entries in the roster and many new members. In addition, the Board has begun updating and expanding the roster using information from the new AEA directory.

---

[5]CSWEP's roster contains information about characteristics such as current employer, educational background, fields of specialization, and number of publications.

To expand the number of employers using the roster to identify potential job candidates, the Board now routinely provides prospective employers with information about the contents of the roster and about the new computer-readable forms in which the data can be obtained. The employers contacted by CSWEP include all those placing announcements in *Job Openings for Economists* or the CSWEP *Newsletter*, as well as the chairs of economics departments.

Another major activity was organizing six sessions for the AEA's 1989 annual meeting —three on gender-related topics and three on theory and applications in industrial organization. For 1990, the nongender-related sessions will focus on science, technology, and productivity. The Board also continued for a second year its practice of sending each person asked by the President-elect to organize a session lists of experienced female economists who specialize in the same fields as the organizer. Preliminary results from the Board's study of participation in sessions at the past seven annual meetings indicate that these reminders were associated with greater representation of women.

Publishing three issues of the CSWEP *Newsletter* absorbed a considerable portion of the Board's resources. The *Newsletter* continued to provide information designed to help young economists succeed in the profession, including calls for papers and notification of sources for research funding. Articles covered topics such as an ex-editor's tips on how to write journal articles that will be accepted and the process for awarding grants used by the Economics Program of the National Science Foundation.

Finally, the Board thanks Joan Haworth, the Committee's Membership Secretary, and her staff, for their outstanding work updating the roster, preparing special mailings, and creating customized lists of female economists from the roster. Three Board members, whose terms expire this year, have also contributed a great deal to the Committee. Marjorie McElroy was the southern regional representative, responsible for organizing sessions, a business meeting, and a reception at each annual meeting of the Southern Economic Association during her term. In addition, she and Cecilia Conrad each co-edited an issue of the *Newsletter*. McElroy, Conrad, and Judith Lave all organized sessions for the AEA's annual meetings and reviewed papers from them for publication. The Board is also grateful to Dennis B. Melby for doing an excellent job in producing the *Newsletter*.

NANCY M. GORDON, *Chair*

Report of the Committee on U.S.–China Exchanges in Economics

In spite of the unfortunate political events in June, China's economics education and economics exchanges with foreign countries have remained essentially unchanged.

On June 19, Chinese participants in the Fourth Biannual Conference of United States-Asia Economic Relations, sponsored by The American Committee on Asian Economic Studies in cooperation with The Pacific Basin Study Program of Columbia University, arrived in New York to deliver their papers as scheduled. In September, nearly all Chinese students planning to come to the United States to pursue their studies in economics and other subjects arrived as planned.

On August 23–24, members of the U.S. Committee on Economics Education and Research in China, including Gregory Chow, Robert Dernberger, D. Gale Johnson, Lawrence Lau, and Dwight Perkins, went to Beijing to meet with the Chinese Committee on Economics Education Exchanges with the United States and colleagues of the State Education Commission to discuss cooperative activities as originally scheduled. At the meeting, it was decided that the main cooperative programs that were to expire in August 1990, including the two economics training centers at the People's University and Fudan University, would continue until August 1992, and that the U.S. committee would help strengthen Ph.D. training at selected Chinese universities by using the above training centers and sponsoring one-year visits of Ph.D. candidates to the United States. Four American faculty members arrived at the two training centers to assume their teaching duties in the fall semester as they had agreed, Moshe Justman and Wing Thye Woo at the People's University, and Charles Hultman and John Huttman at Fudan University. Anyone interested in teaching at these two training centers may contact Dr. Todd Johnson, Executive Director, Committee on Economics Education and Research in China, National Academy of Sciences, 2101 Constitution Avenue, Washington, D.C. 20418.

On November 1–3, eight American economists went to Shanghai to participate in a conference with Chinese colleagues to discuss drafts of new economics texts being written for use in Chinese colleges and universities. In 1987, the Chinese State Education Commission had accepted the recommendation of an international panel of consultants of the World Bank to include as core courses for economics majors the subjects of micro- and macroeconomics, development, comparative economic systems, statistics, computer applications, public finance, money and banking, international trade and finance, and accounting along with Marxian economics. Committees of Chinese economists were appointed to prepare texts for these courses. Texts for the first four subjects listed above were discussed in the November 1–3 conference that had originally been scheduled in July. The American participants were impressed by the quality of the texts prepared and by the openmindedness of their Chinese colleagues in the discussion. This is an activity of the Chinese University Development Project II sponsored by the World Bank. Programs to train teachers for the core subjects were being planned by officials of the State Education Commission. It will take some years before the core courses can be successfully implemented at the large number of Chinese colleges and universities.

On August 27, three American members attended a meeting of the Board of Trustees of Lingnan (University) College of Zhongshan University in Canton. At the meeting, the introduction of micro- and macroeconomics, economic development, and international trade into the economics curriculum was being planned. Visiting professors from the United States, Great Britain, and Hong Kong were to be invited to teach these subjects in the 1989–90 academic year.

It is difficult to forecast the ultimate impact of the political events of June. However, continued modernization of economics education and continued economics exchanges with the United States and the outside world appear to have survived and may even assert some influence in the future development of China.

GREGORY C. CHOW, *Chair*

Report of the *Ad Hoc* Committee on the *Journal of Economic Perspectives*

In the spring of 1988 the Executive Committee of the American Economic Association formed a five-member *Ad Hoc* Committee to study the newly founded *Journal of Economic Perspectives*. This group was asked to report to the Executive Committee on two matters. The first was the reception accorded to the new journal by the membership of the association, with the *Ad Hoc* Committee invited to offer any comments of its own on the quality of the journal and its contents. The second and more explicit task assigned to the committee was to review the arrangements for the financing of the journal, considering whether it should continue to be provided automatically to all members of the AEA, with the costs covered largely out of membership dues, or whether, instead, the Association should undertake some form of unbundling of its journal subscriptions, offering members the option of taking some proper subset of the three publications along with some appropriate reduction in membership fee.

In fairly extensive correspondence, the *Ad Hoc* Committee undertook to conduct a survey of a sample of AEA members. This was, in fact, carried out and the results analyzed with the very competent and cheerful help of the personnel of the AEA's Nashville office. However, before the survey could be conducted, it was necessary to plan the design of the questionnaire with considerable care since, as we know, the nature and coverage of the results, as well as the responses themselves, can be influenced heavily by the included questions and their phrasing. In particular, it was necessary to study in advance what pricing options should be proposed for evaluation by the respondents to the survey, that is, what form of unbundling should be considered, given the great desirability of avoiding excessive length and excessive complexity in the questionnaire.

In designing the questionnaire the *Ad Hoc* Committee had the full cooperation of the editors of *JEP*, who suggested the addition of several searching questions likely to provide information helpful in editing the journal. These questions were, of course, included. The survey also invited respondents to send in any pertinent comments they wanted to make and a notice in the *JEP* extended a similar invitation to all other readers of that journal.

The analysis of the survey results and the comments of the members are now in hand and, after an exchange of correspondence, our committee has prepared its report on the basis of these materials.

Conclusions and Recommendations

In summary, we have reached the following conclusions:

a) Most members of the AEA like the *JEP*. Its average approval rating is not significantly different from that accorded the *AER* or the *JEL*.

b) However, each of these journals is disliked intensely by some segment of the members. There are particularly strong feelings among what may or may not be a small group of members against the *AER* which some members consider to have been captured by an insider group drawn from "the establishment," and whose contents are believed by this group to be composed heavily of the esoterica that interest only the establishment readers, who are not the majority of the AEA members.

c) Some members strongly resent the *JEP* policy against unsolicited submissions, asserting that it invites assignment of the publication opportunities by clique, and that it slants the contents of the journal even more heavily than usual toward the preferences of the editors, in contradistinction to those of the bulk of the AEA members.

d) Most members like the individual features of the *JEP* and approve of its contents generally.

e) A majority of the respondents, which was hardly overwhelming (58 percent in favor, 42 percent against) favor *some* unbundling of journal subscriptions. A very small group of members favored drastic unbundling.

f) Fifty-seven percent of the respondents said that if unbundling were permitted, they would nevertheless continue to subscribe to all three AEA journals.

g) The write-in comments showed no consistent pattern except, perhaps, for fairly frequent questioning of the *JEP*'s policy on unsolicited manuscripts.

Accordingly, the *Ad Hoc* Committee recommends:

1) That the Executive Committee of the AEA reaffirm its support of the *JEP*.

2) That the editors of the *JEP* be urged to give serious consideration to the opinions of the members of the AEA as expressed in their letters and in the results of the survey.

3) In particular, the editors of the *JEP* should be urged to continue to monitor and assess their policy on unsolicited submissions. Here it should be emphasized that there are good reasons for continuation of the current policy and for full support by the Executive Committee of the editors of the *JEP* in their apparent decision to continue the current course.

On balance, the *Ad Hoc* Committee also favors continuation of the current policy, under which articles are almost exclusively commissioned by the editors, since a decision to discontinue this arrangement may well amount to a decision to revise the character of the journal itself. Given the evidence of general member satisfaction with the contents of the journal, such a change appears to us to be inappropriate. If the Executive Committee accepts this recommendation, this should be considered tantamount to a strong and explicit statement by it to this effect, one designed to strengthen the position of the editors of the journal.

4) Since the editor of the *JEP* (or the delegates of the editor) has such a major influence on the articles that will be printed in the journal, it may be appropriate to change *JEP* editors somewhat more often than the other journals do, and perhaps, in addition, to have the selection process for editors be somewhat more open. If the *JEP* editor were to change, say, every three years, so that a variety of editors were chosen from different geographic areas, different types of academic institutions, and different economic specialties, the number of complaints about the openness of the journal might be substantially reduced.

Alternatively, it may be possible to achieve the same objectives by appointment of an editorial board whose membership is heterogeneous, and whose opinions are given considerable weight in the running of the journal. This latter option is the one favored by the current editors of the journal, who report that they are in fact operating in accord with it. The *Ad Hoc* Committee sees no reason to recommend the imposition of a change in this policy.

5) The Executive Committee should *consider* undertaking a three-year unbundling experiment. However, in considering this matter it should be recognized that

a) Once inaugurated, such a step may prove difficult to reverse;

b) That the financial savings to members who drop a journal will have to be low since the marginal cost of a copy of a journal is very small;

c) That unbundling should not favor any one of the three Association journals, making any of them into "more-official" publications of the AEA than others. In other words, if unbundling is decided upon, members should be offered the opportunity to drop *any one* journal they select, with a corresponding reduction in annual dues.

d) In view of the close vote on the unbundling subject in general, and the absence of any substantial number of advocates of any radical unbundling measure, no drastic unbundling arrangement seems to be justified.

e) If the Executive Committee should decide to undertake the unbundling experiment, the results in terms of membership satisfaction and financial consequences should be reviewed after three years, possibly with the aid of a follow-up survey.

The *Ad Hoc* Committee favors, rather weakly, the undertaking of the unbundling experiment, but recognizes the validity of the counterarguments, which have already been summarized here.

6) Any unbundling should proceed hand-in-hand with an offering of the journals to readers outside the AEA. If members need

not receive all the journals, it makes some sense for nonmembers to have easier means to receive at least one of the journals. This is already happening to a limited extent; for example, members of the European Economic Association now have an option to receive *JEL*, if they so desire, in addition to the *European Economic Review*.

We turn now to a description of some of the pertinent details of the *Ad Hoc* Committee's study.

The Nature of the Survey and its General Results

The membership survey was carried out by mailing copies of the questionnaire to 1000 members of the AEA, who were selected alphabetically from the total membership list, to yield a rough approximation to a random sample. One month after the initial mailing, 376 responses had been received, and a follow-up solicitation yielded an additional 148 respondents, constituting altogether a 52 percent response rate.

The questionnaire dealt with the intensity with which recipients of *JEP* make use of the journal, their overall satisfaction with its contents, their evaluation of its various features, and their views on the methods of financing of the journal. A copy of the questionnaire is appended to this report, along with a copy of the analysis of the survey results that was prepared for our committee by Christopher Schenzler of Vanderbilt University.

In describing the results of the survey, the discussion here of readers' views on the quality and value of the journal, and of its contents will be rather brief, since the Executive Committee presumably will take little detailed action on these matters, though the results in these areas will presumably be of considerable interest to the journal's editors. In contrast, our discussion of the survey results relating to the financing of the journals of the AEA will be rather more detailed.

On the first of these two broad subjects, the survey indicated that the majority (90 percent) of respondents read at least one or two articles in every issue of *JEP*, and that they read the regular features fairly consistently. They use the journal, above all, to keep up with developments in subfields of economics other than their own. They generally consider the manner of presentation of the materials to be about right (again, about 90 percent) and consider the journal to be reasonably valuable overall (61 percent), with another 20 percent considering the journal to be extremely valuable. Thus, the survey surely suggests that *JEP* has generally been well received by the membership of the Association, and there appears to be every reason to favor its continuation and continued support by the Executive Committee of AEA.

Few changes in direction were suggested by the results of the survey. Many of the respondents indicated that they would like more discussions of current policy issues and a broader range of articles in each issue of the journal. And, as already noted, quite a few respondents took exception to the journal's policy on unsolicited manuscripts. Aside from the expected number of disgruntled respondents who had special complaints, this seems to constitute the bulk of the calls for modification in the contents or the workings of *JEP*.

On the Financing of the AEA Journals

Some preliminary remarks are necessary to explain the wording of the questions on the financing of the AEA journals. The basic goal underlying the questions was to offer a choice between continuation of the current arrangement, under which all three journals are supplied to all members, with no charge beyond the membership dues, and a new unbundling arrangement, under which every member would be offered the option of dropping any one of the three journals, with a corresponding reduction in membership fee. Elton Hinshaw was first asked whether the second of these arrangements would impose an impossible burden upon the administrators of the AEA, and we were assured that it would probably cause no major difficulties.

The next issue that had to be dealt with was the magnitude of the reduction in membership fees to be offered in the second option. Here, a somewhat disturbing asymmetry intruded itself. To see why this is so, consider a case in which an association inaugurates a new journal that is to be financed without profit or loss via a nondiscriminatory rise in membership fees. In that case, the rise in membership fees must equal the total incremental cost of producing and distributing the new journal, divided by the numbers of members of the association. That is, the rise in membership fee must equal the *average incremental cost* of the journal. Next, suppose a member is given the choice of dropping one of the journals. The zero-profit objective then requires a reduction in membership fee to any member who decides to drop the journal—a reduction equal to the *marginal cost* of the journal. Since in journal publication marginal cost is well below incremental cost, it follows that the saving that can be offered the member who drops one journal will be significantly smaller than the initial rise in dues needed to finance the new journal.

To offer respondents a realistic depiction of the options, Elton Hinshaw was asked to provide estimates of the marginal and incremental costs of the AEA journals, and on the basis of figures that he supplied, our financial questions were formulated.

Specifically, respondents were asked the following question:

> Two of the proposed methods to pay the cost of the new journal are:
>
> *Option A*: Raise annual dues about $12 (approx. avg. incremental cost) per member and provide all three journals to all members.
>
> *Option B*: Permit members to request that any one of the three AEA journals no longer be sent to them. Members who selected this option would receive a rebate of $6 (approx. marginal cost).
>
> 1. I prefer: (a) option A; (b) option B ____.

Some 93 percent of the respondents answered this question. Of those who answered, some 42 percent favored continuation of the current financing arrangement, while 58 percent, a clear majority, but not an overwhelming one, voted for the unbundling option.

This result is given more substance by the follow-up question that asked respondents whether, if the unbundling regime were instituted, they would continue to subscribe to all three journals, or, if they would not, which journal they would drop. Here, the response rate was 86 percent. Of the answering group, 57 percent indicated that they would continue to subscribe to all three journals. The remainder were almost equally divided in the choice of journal to be dropped —15 percent expected to drop the *AER*, 14.2 percent the *JEL*, and 13.8 percent the *JEP*.

The important point to be noted is that, despite the 58 percent of the respondents who favored the unbundling option, almost exactly the same percent of the respondents indicated an intention *not* to take advantage of the option if it is offered. This result goes far to explain the guarded recommendation of our committee on the subject of unbundling. It certainly helps to explain our outright rejection of the few proposals for a more radical unbundling arrangement, for example, the suggestion that the AEA drop all three journals from their status as official Association publications, offer none of them to the AEA members in return for membership dues, and set all three journals adrift to compete for subscribers as best they could, on the open market.

Our Thanks

In concluding, we must not neglect to express our thanks to Christopher Schenzler for carrying out the analysis of the survey so ably and so promptly, and to Timothy Taylor, the managing editor of *JEP*, for his helpful suggestions and for communicating the views of that journal's other editors to us. Above all, we must indicate our gratitude to Elton Hinshaw who met all our needs quickly, effectively and promptly. His help was critical for completion of the work of our committee—Leonid Hurwicz, Michael S. McPherson, Albert Rees, and Robert M. Solow.

WILLIAM J. BAUMOL, *Chair*

Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research studies a wide variety of economic issues. Much of this research is conducted by economists working individually, but a large part is organized into special projects. To disseminate the results of this research, during 1989 the NBER issued over 400 working papers, published 10 books, 2 NBER journals, prepared 4 special issues of other journals, and circulated the monthly *Digest* and the quarterly *Reporter*. In addition, the NBER held numerous conferences and workshops, including the 6-week-long Summer Institute. Finally, the NBER hosted 4 Olin Fellows who spent the year conducting empirical research on topics of their own choosing.

Programs. The NBER's ongoing programs generally meet twice during the academic year and once, for a longer period, during the Summer Institute. Bureau programs (with directors in parentheses) are Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), and Productivity (Zvi Griliches).

Projects. The NBER's projects generally bring a dozen or more researchers together to work on a common topic. The projects' findings are usually distributed initially as NBER working papers, and final versions are then published as Bureau books. During 1989, the following projects (organizer in parentheses) were underway: Economic Crisis (Martin Feldstein), Economic Growth (Robert Barro and Paul Romer), Economics of Higher Education (Charles Clotfelter and Michael Rothschild), Financial Time-Series (John Campbell and Angelo Molino), Information and Corporate Finance (Glenn Hubbard), International Taxation (Assaf Razin and Joel Slemrod), Macroeconomic History (N. Gregory Mankiw and Christina Romer), Political Economy (Alberto Alesina, Moe Fiorina, and Roger Noll), Savings (Douglas Bernheim and John Shoven), Stock Market Volatility (Sanford Grossman), Strategic Trade (Paul Krugman and Alasdair Smith), and the United States and Japan in the 1990s (Paul Krugman).

Other Conferences. In addition to the conferences and workshops conducted as part of NBER projects and programs, the NBER also holds larger conferences open to individuals in government and academe. During 1989, these included a session of the Conference on Income and Wealth on the Measuring International Economic Transactions, organized by Peter Hooper and J. David Richardson; a Universities Research Conference on Social Insurance, organized by Douglas Bernheim; and a second URC on Labor Markets in the 1990s, organized by Robert Topel. Procedures and deadlines for submitting papers to these conferences are posted in economics departments and announced in the NBER *Reporter*.

Summer Institute. During the summer of 1989, 603 economists from 158 different universities and research organizations attended the Summer Institute, including 215 participants who came for the first time. Participants stayed for periods ranging from 4 days to 2 months. There were 253 papers presented in 31 separate sessions. A catalogue of these papers, including authors and their addresses, is available from the NBER on request.

Working Papers. The NBER working papers are circulated to economics department libraries around the world, and are available for $2 per title from the NBER. Subscriptions to the entire series are available for $300 per year inside the United States and Canada, $500 other foreign.

Books. The 10 books published by the NBER during 1989 included 3 volumes on Developing Country Debt. Jeffrey Sachs edited this series, and received assistance from Susan Collins in editing a volume of studies of Indonesia, Korea, the Philippines and Turkey. Michael Bordo edited a collection of essays on Money and International Finance that were written in honor of Anna Schartz. David Wise edited papers on Economics of Aging; Robert Feenstra edited a volume on Trade Policies and Competitiveness; and Robert Lipsey and Helen Tice edited a volume on the Measurement of Saving, Investment, and Wealth. In addition, Claudia Goldin wrote an economic history of American Women, and Charles Clotfelter and Philip Cook wrote a monograph on state lotteries.

Journals. The NBER publishes two journals, which appear once each month. Olivier Blanchard and Stanley Fischer edited the *Macroeconomics Annual* in 1989, and Lawrence Summers edited *Tax Policy and the Economy.* Subscriptions for these two journals can be ordered from MIT Press.

In addition, the NBER sponsored special issues of several journals. Robert Gordon and Georges de Menil edited papers on international macroeconomics for the June issue of the *European Economic Review.* James Poterba edited papers on residential real estate and capital formation for the May issue of *Regional Science and Urban Economics.* Edmar Bacha and Sebastian Edwards edited papers for publication in the *Journal of Development Economics.* Robert Baldwin and Alan Winters organized articles on international trade policy for the *Weltwirtschaftliches Archiv.*

Periodicals. The monthly NBER *Digest* provides brief summaries of working papers of general interest. The *Reporter* contains longer summaries of recent research, abstracts of all working papers, announcements of the publication of NBER books, a survey of economic forecasters, and a calendar of forthcoming conferences. Both these periodicals are available on request.

Olin Fellows. The 4 Olin Fellows for the 1989–90 academic year are Alberto Alesina, Alan Krueger, Karen Lewis, and David Scharfstein. These fellowships are open to young empirical economists at any academic institution. Application deadlines are announced in the *Reporter.*

During 1989, Martin Feldstein continued as President of the NBER and Geoffrey Carliner continued as Executive Director. Further information on NBER activities is available in the *Reporter* or from Geoffrey Carliner, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138.

DAVID KENDRICK, *Representative*

Report of the Representative to the American Association for the Advancement of Science

The American Association for the Advancement of Science is a federation of scientific organizations, as well as an association of over 135,000 individual members. Its objectives are "to further the work of scientists, to facilitate cooperation among them, to foster scientific freedom and responsibility, to improve the effectiveness of science in the promotion of human welfare, and to increase public understanding and appreciation of the importance and promise of the methods of science in human progress."

The discipline of Economics is grouped with Political Science, Sociology, Geography, and some related social sciences in Section K, one of the twenty-one divisions within the AAAS. Most economists list Section K as their primary affiliation, and some hold dual or multiple membership in other sections. Overall, the four social science sections comprise just under 10 percent of the AAAS-affiliated membership.

The AAAS offers economists a unique opportunity to interact with the scientific community at large. This includes the following channels: the AAAS annual meeting, joint sessions with other organizations at their professional meetings, specialty conferences, *Science*, and AAAS-sponsored books. In addition, the organization is active through the work of its offices and standing committees on Science and Technology Education, Public Sector Programs, Opportunities in Science, International Science, Scientific Freedom and Responsibility, and Science, Arms Control and National Security.

A joint AEA/AAAS session was held at the 1989 ASSA meetings on "Research Issues in the Economics of Climate Change." The session was chaired by Mancur Olson and included papers by Faye Duchin, Dale Jorgenson and Peter Wilcoxen, and Gary Yohe.

Ten symposia and four technical sessions at the 1989 AAAS annual meeting, held in San Francisco in January, were devoted to "Economics and Competitiveness." Participants included James Dorian, Irwin Feller, Edward Malecki, Karen Polenske, Dorothy Rice, Sherman Robinson, and John Tilton. Economists such as William Brock, Charles Plott, Harry Richardson, and Vernon Smith participated in sessions listed under other disciplinary headings.

The visibility of the economics profession at the AAAS meeting has been an issue for several years. Only a limited number of sessions is devoted to Economics because the Head of the AAAS Meeting Office believes attendance at these sessions is inadequate. At the same time, the incentive for economists to attend the annual AAAS meeting is diminished if there are few sessions by their colleagues. Also, the economics profession finds it sessions placed in other disciplinary categories in order to broaden the audience base. Sessions involving prominent economists have been listed under such headings as Sociology and Geosciences, thus making them difficult to identify on the Annual Meeting Program. Perhaps even more problematic is the fact that other disciplines are grouped with Economics, or that sessions involve discussions of popularized topics, rather than economic science, in a further attempt by the AAAS Meeting Office to attract a general, and, thus, possibly, larger audience. The AAAS Meeting in Washington, D.C., in February 1991, provides an excellent opportunity to correct this situation. Every attempt will be made to offer several sessions by prominent economists, in addition to the many quality sessions offered by the scientific community at large. The AEA members in the Washington Metro area are especially encouraged to attend this meeting to help establish a record of a sufficient attendance base to ensure self-determination for the discipline at future AAAS conferences.

Continuing the trend of the past two years, several economists were commissioned to write papers for *Science*. Authors of commissioned or contributed papers included: William Baumol, Michael Hurd, Edward Kane, Burton Malkiel, and Burton Weisbrod. Also, Vernon Smith was appointed to the Board of Reviewing Editors of *Science*.

Several economists are officers of Section K. They include: Mancur Olson (Chair), Robert Fogel (Section Committee Member), Richard Cyert, Lester Lave, and T. Paul Schultz (Electorate Nominating Committee Members).

Further information about the AAAS is available from Marge White, 1333 H Street, N.W., Washington, D.C. 20005. AEA members are also encouraged to contact members of the AEA/AAAS Liaison Committee: Roger Bolton, Gardner Brown, Faye Duchin, Adam Rose, Vernon Smith and Robert Solow, about AAAS activities.

ADAM ROSE, *Representative*

Policy and Advisory Board of the Economics Institute

During more than 30 years, the Economics Institute in Boulder, Colorado has provided preparatory training for foreign students coming to North America to undertake graduate study in economics, agricultural economics, business and public affairs. During that period, thousands of students have attended the Institute. The Institute experience has shortened the period needed to complete graduate degrees and improved the quality of graduate education for thousands of alumni. For others, it has meant the difference between success and failure in a once-in-a lifetime opportunity for foreign students.

During its entire existence, the Economics Institute has been under the same Director, Wyn Owen. He founded the Institute, defined its mission, nourished its growth, managed an increasingly complex institution and tended to thousands of details without ever losing sight of its basic goals. Wyn Owen retired as the Economics Institute's first and only Director at the end of December 1989. His Board salutes him.

After a thorough search, the Board appointed Charles Becker to be the new Director effective January 1, 1990. Charles Becker received the Ph.D. in economics at Princeton University, taught at Vanderbilt University for several years, and has been Deputy Director for the Academic Program at the Institute for two years. The Board is pleased to have a new Director of Becker's quality, achievements and dedication.

Institute enrollment in 1989 continued to be impacted by foreign debt crises in many Third World countries. However, enrollment declines reported in recent years appear to have stopped, and we look forward to resumed growth. Although cutbacks are painful, the Institute has continued to provide its traditional high quality preparatory training.

Programs and institutions interested in further information about the Institute should contact Director, Economics Institute, 1030 13th Street, Boulder, CO 80302 (telephone 303 492-3000; Telex 450385 ECONINST BDR).

EDWIN S. MILLS, *Chairman*

GAUSS THE CHOICE OF THE EXPERTS

“The speediest and most versatile of the mathematical programming languages.”

Barry Simon - PC Magazine

PC MAGAZINE EDITOR'S CHOICE MARCH 14, 1989

You get more work done when you use the right language.

Fast and Flexible

Now you can discover why scientists, engineers and statisticians agree that GAUSS is the language of choice when you want the most in speed and flexibility from your PC. The built-in matrix routines are written in assembly language and are faster than even the fastest C or Fortran compilers.

Multiple Regression

```
xxi = inv(x'x);
b = xxi*(x'y);
sse = (y-x*b)'(y-x*b)/(n-rows(b));
stderr = sqrt(diag(sse*xxi));
tvalue = b / stderr;
pvalue = cdftc(tvalue,rows(x));
```

It's not just execution speed that makes GAUSS the better choice. GAUSS is a matrix language, using familiar syntax for expressing mathematical relationships. The result is you will shorten your development time and write less code. GAUSS's programming environment has all the tools you will need to write large programs, including an on-line debugger and an on-line help system that can handle the functions you write.

Expandable

Create your own libraries of functions or call FORTRAN, C, or assembly language routines.

Publication Quality Graphics

The GAUSS system includes 2-D and 3-D publication quality graphics with an underlying resolution of 4096 x 3120.

Applications Modules

The optional SimGauss module simulates nonlinear differential equations, state-space and discrete systems. It's powerful, flexible and supports state/parameter vectors for increased speed and ease. Modules are also available for numerical optimization, solving simultaneous nonlinear equations, generating basic descriptive statistics, maximum likelihood, logit, probit, loglinear analysis, and others.

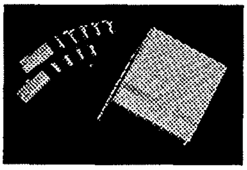
Economical

For just $495 you get the basic system including a library of mathematical functions and publication quality graphics.

Guaranteed

If you don't agree that GAUSS is the speediest and most versatile of the mathematical programming languages, simply return it for a full refund. No questions asked. Call (206) 631-6679 to order.

GAUSS–faster, easier, smarter... guaranteed.



System includes 700 page manual and 11 program and application disks for IBM PC-XT-PS/2 & compatibles with math coprocessor. 386 protected mode version also available.

GAUSS™

BY APTECH SYSTEMS, INC.
26250 - 196th Place SE, Kent Washington 98042
Phone (206) 631-6679 FAX (206) 630-1220
CALL OR WRITE FOR FREE BROCHURE

SPRINGER ON ECONOMICS

☐ Topics in Advanced Econometrics: Probability Foundations

by **P. J. Dhrymes**,
Columbia University, New York, NY

P. J. Dhrymes treats econometrics with the mathematical incisiveness necessary to reach the fundamental nature of a statistical event. For the professional without the benefit of formal training in mathematics **Topics in Advanced Econometrics** is an essential book for understanding contemporary econometric theory. **Dhrymes** restrains himself from the type of generalizations that would lead the reader to the naive solution to the coin toss problem. Instead he exposes the underlying suppositions of econometric topics.

Contents: Mathematical Foundations. Foundations of Probability. Convergence of Sequence I. Convergence of Sequence II. Dependent Sequences.

1989/379 pp./hardcover $39.95/ISBN 97178-5

☐ Advanced Transport and Spatial Systems Models: Applications to Korea

by **T. J. Kim**,
University of Illinois at Urbana-Champaign

The main thesis of this book is that the characteristics of socio-economic activity that best describe mixed economic systems are (1) interaction between the public and private sectors and (2) the derived nature of transportation demand.

Contents: National Development Plans and Planning Issues in Korea. The transportation Systems in Korea. National Transporation Development Model. Implementation of the Linear Programming Model: Experience from a Korean Application. Combine Input-Output and Commodity Flow Model: A Nonlinear Programming Formulation. Model Applications: Formulating Highway Investment Strategies. Bilevel Programming. Bilevel Transportation Network Design Models and Solution Algorithms. Model Application I: Measurement of Congestion Functions for Korean Highways. Model Application II: Evaluating Alternative Highway Investment Strategies. Integrated National Transport Planning Model for Mixed Economic Systems: Bilevel Programming Approach.

1990/255 pp./27 illus./hardcover $39.00 (tent.)/ ISBN 97277-3

☐ The Collected Economics Articles of Harold Hotelling

Ed. by **A. Darnell**,
University of Durham, U.K.

Here are nearly all of the published articles in economics of one of the most successful twentieth-century pioneers of mathematical economics. Many economists today find inspiration in Hotelling's pathbreaking papers, and this collection should provide even further inspiration for research.

1989/178 pp./10 illus./hardcover $49.80/ ISBN 97011-8

☐ Business Cycle Theory: A Survey of Methods and Concepts

2nd Revised & Enlarged Ed.,
by **G. Gabisch & H.-W. Lorenz**,
both of University of Göttingen, FRG
Universitext

Presents a comprehensive survey of modern business cycle theory, encompassing shock-dependent and shock-independent business cycle models. This revised and enlarged edition includes a separate chapter devoted to new developments in dynamical economics.

1989/248 pp./75 illus./softcover $34.00/ ISBN 51059-1

To order

call 1-800-SPRINGER or 201-348-4033 in NJ; or send a copy of this ad with check or AmEx, MC, or VISA number and expiration date, plus $2.50 for shipping, to: **Springer-Verlag New York, Inc.,** Attn: John DiStefano, Dept. S 467, 175 Fifth Avenue, New York, NY 10010.

Name____________________________

Address__________________________

City/State/Zip_______________________

Circle one AmEx MC VISA

Card No.________________ Exp. Date________

Signature___________________________

Springer-Verlag
New York Berlin Heidelberg Vienna
London Paris Tokyo Hong Kong